Understanding Bias in Using Machine Learning for Loan and Credit Guidance.

by Sandip Biswas
Student ID: 12689982

University of Essex

Supervisors:
Dr. Samuel Danso
Douglas J. Millward

Abstract

Machine learning (ML) has become ubiquitous in decision-making, including loan approvals and credit guidance. However, ML models risk amplifying societal biases, leading to discriminatory outcomes (Barocas & Selbst, 2016). This dissertation critically examines bias in ML-driven loan and credit guidance systems, analysing its forms, real-world consequences, and mitigation strategies. Bias can manifest as historical bias (from past discriminatory practices), representation bias (from unrepresentative data), measurement bias (from data collection errors), or proxy bias (from features correlated with protected attributes like race or gender). Such biases perpetuate socioeconomic inequalities and violate fair lending regulations. This research will provide a comprehensive literature review on bias in ML lending, empirical analysis of real-world or simulated loan datasets, evaluation of mitigation techniques, and recommendations for best practices to promote fairer lending outcomes. It will emphasise the ethical considerations for building and responsibly using ML models within the financial lending sector.

Acknowledgements

I sincerely thank my supervisors, Dr. Samuel Danso and Douglas J. Millward. Their patience, guidance, and insightful advice have been instrumental in completing this dissertation.

Furthermore, I am sincerely grateful for the unwavering support of my family and close friends. Their continued encouragement has provided essential motivation throughout my research endeavours.

Table of Contents

1. Introduction

Machine learning (ML) algorithms have become omnipresent facilitators of decision-making across various domains. With its perceived objectivity and power to process large amounts of data, ML increasingly assists in financial decision-making processes, specifically in loan approvals and credit guidance. However, growing evidence reveals that ML models can embed and even amplify existing societal biases, leading to discriminatory and unjust outcomes (Barocas & Selbst, 2016). Understanding and addressing bias within ML models designed for loan and credit guidance has become paramount for ensuring fairness, ethical practices, and consumer protection.

This dissertation investigates the multifaceted nature of bias within ML systems used in financial lending. It critically analyses how bias manifests at different stages of the ML pipeline, delves into the potential socioeconomic implications of biased models, and examines strategies for mitigating bias to enhance the fairness of lending decisions.

1.1. Background and Motivation

Traditionally, financial institutions relied on human judgment and limited data points for loan approvals and credit assessments. This process was inherently subjective and prone to human biases (Bertrand & Mullainathan, 2004). With ML, the hope was to replace this with standardised, data-driven decision-making. However, this technological evolution is not immune to bias.

ML algorithms function by learning patterns from historical data. If this data reflects past discriminatory practices or systemic social inequalities, the ML model risks perpetuating and amplifying those biases (Mehrabi et al., 2021). For instance, an ML model trained on data containing historical gender or racial disparities may be less likely to approve loans for women or minority applicants, even if they have equally strong credit profiles. Such embedded bias within ML-powered credit guidance systems can exacerbate socioeconomic inequalities.

Aside from the ethical concerns, biased ML models pose significant risks for financial institutions. Regulatory bodies are becoming increasingly attentive to algorithmic fairness, and institutions may face legal and reputational consequences for discriminatory practices resulting from biased algorithms (Lepri et al., 2018). Moreover, biased models could only allow credit to deserving borrowers, curtailing economic opportunities and stifling potential growth.

1.2. Forms of Bias in ML-Based Lending

Bias can enter ML-driven lending systems in subtle and complex ways. The main types of bias include:

- **Historical Bias:** This occurs when ML models' training data reflects past discriminatory practices. For example, a model trained on data from an era when lending was less accessible to minorities could reproduce those discrimination patterns (Kleinberg et al., 2018).

- **Representation Bias:** This form of bias arises if the data is not fully representative of the population it will be used to evaluate. A loan model trained primarily on data from high-income applicants may lack the ability to assess low-income borrowers, leading to biased outcomes accurately (Hardt et al., 2016).

- **Measurement Bias:** This occurs due to errors or inconsistencies in measuring or collecting features (input variables). For example, using self-reported income data in a loan model could be inaccurate and introduce bias if some demographics are more likely to understate or overstate their income (Veale & Binns, 2017).

- **Proxy Bias:** This bias happens when seemingly neutral features used in an ML model become proxies for protected attributes like race, gender, or age. For instance, using Post Codes as a feature in a lending model can create a proxy bias if specific Post Codes are strongly correlated with racial demographics (Bartlett et al., 2022).

1.3. Real-World Examples

The bias in algorithmic lending is not a theoretical concern but a well-documented issue. Some notable examples include:

- **Apple Card:** In 2019, the Apple Card credit service was scrutinised for allegedly offering women lower credit limits than men, even with comparable financial profiles (Nellis, 2019). This raised concerns about potential gender bias embedded within the algorithm.

- **Amazon Hiring Tool:** Reports emerged that an AI-driven recruitment tool developed by Amazon exhibited bias against women. The system had reportedly been trained on resumes from predominantly male candidates, causing it to downgrade resumes containing words associated with women (Dastin, 2018).

- **COMPAS Software:** Widely used in the US correctional system, the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm assesses recidivism risk. Studies have found it to be biased against black defendants, incorrectly flagging them as high-risk more frequently than white defendants (Angwin et al., 2016).

- **FICO Credit Scores:** While not strictly machine learning-based, FICO scores are widely used in lending decisions and can be influenced by biased features. Research suggests racial disparities exist in FICO scores, potentially due to underlying historical and socioeconomic factors. (Bartlett et al., 2022). Investigating the specific features of credit scoring models that perpetuate bias is critical to understanding bias in lending.

- **LendingClub Discrimination Case:** In 2016, the peer-to-peer lending platform LendingClub came under regulatory scrutiny for charging higher interest rates to minority borrowers, even when other risk factors were similar. This case highlighted how traditional lending platforms, even those with a tech-driven approach, might inadvertently embed bias in their pricing algorithms. (Bloomberg, 2016).

- **Biased Creditworthiness in Underbanked Communities:** ML models trained on conventional credit data may need help accurately assessing the creditworthiness of underbanked individuals or those with thin credit files. This can result from using features disproportionately available to traditionally well-banked populations, leading to bias against those who lack extensive credit history (Wang, 2021).

- **ZestFinance Alternative Data Usage:** ZestFinance, a fintech lender, employs alternative data sources (such as social media activity or browsing patterns) in their ML models. While promising for assessing unconventional borrowers, such data sources risk introducing new forms of bias. There is a need to evaluate unintended correlations with protected characteristics within such alternative data (The Wall Street Journal, 2015).

1.4. Impact of Algorithmic Bias on Lending

The presence of bias within ML-based credit guidance systems has wide-ranging implications:

- **Perpetuating Systemic Inequality:** Biased ML models can further entrench socioeconomic disadvantages by denying qualified individuals from marginalised groups access to financial opportunities (Citron & Pasquale, 2014).

- **Diminishing Trust:** As instances of algorithmic bias gain public attention, trust in financial institutions and the fairness of their lending processes is likely to erode (Polonski, 2021).

- **Legal and Regulatory Ramifications:** Financial institutions operating with discriminatory algorithms may face legal action for violating fair lending regulations and anti-discrimination laws (Pasquale, 2015).

- **Reputational Damage:** Adverse publicity stemming from cases of algorithmic bias can cause long-term harm to financial institutions' reputations and brand image.

1.5. The Need for Fairness and Bias Mitigation

Ensuring fairness in ML-powered lending is a complex challenge and a moral and business imperative. To combat bias, a multi-pronged approach is required, including:

- **Algorithmic Auditing:** Regularly auditing ML models to identify and quantify potential bias (Veale et al., 2018).

- **Bias Mitigation Techniques:** To mitigate detected biases, various techniques, such as pre-processing data, in-processing adjustments to the algorithm design, or post-processing of outputs, are employed (Hajian et al., 2016).

- **Explainability:** Developing techniques for understanding how ML models make decisions, allowing for better identification of bias sources (Doshi-Velez & Kim, 2017).

- **Robust Governance:** Implementing strong governance structures within financial institutions to oversee the responsible deployment and ongoing monitoring of ML systems for lending (Heller et al., 2022).

1.6. Roadmap for Dissertation Research

This dissertation aims to critically examine the issue of bias in ML-driven loan and credit guidance systems. The research objectives are as follows:

1. **Systematic Survey:** Conduct a comprehensive literature review to critically analyse the types of bias present in ML-based lending, frameworks developed to assess bias, and the efficacy of bias mitigation techniques.

2. **Empirical Analysis:** Design and implement an empirical study analysing bias potential in real-world or simulated loan datasets. Develop suitable bias metrics to quantify the extent and nature of bias.

3. **Evaluation of Mitigation Strategies:** Evaluate the effectiveness of various bias mitigation techniques in the context of lending algorithms. Assess their impact on both fairness and model performance.

4. **Recommendations:** Develop actionable recommendations for financial institutions outlining best practices for building, auditing, and mitigating bias in ML-driven credit guidance systems, ultimately promoting fairer lending.

1.7. Research Problem Statement

Machine learning algorithms rapidly permeate financial lending, promising efficiency and standardisation in loan approvals and credit guidance. However, mounting evidence suggests these algorithms are susceptible to embedding and perpetuating societal biases, leading to discriminatory outcomes and severe ethical concerns (Barocas & Selbst, 2016; Mehrabi et al., 2021). The bias in algorithmic lending decisions undermines the moral principles of fairness and the legal frameworks governing access to financial opportunities.

1.8. Importance of Ethical Analysis

A thorough ethical analysis of ML bias in the lending space is critical for several reasons:

- **Fairness and Equity:** Algorithmic bias can unjustly deny marginalised individuals loans and credit, perpetuating economic disadvantage cycles (Citron & Pasquale, 2014). An ethical framework helps ensure ML systems align with principles of distributive justice.

- **Consumer Trust:** As instances of algorithmic bias become more widely known, consumers may become increasingly distrustful of financial institutions and their decision-making processes (Polonski, 2021). Ethical analysis helps build trust by demonstrating a commitment to transparency and accountability.

- **Regulatory Compliance:** Financial institutions are subject to fair lending regulations like the US Equal Credit Opportunity Act (ECOA). Ignoring algorithmic bias could lead to legal repercussions (Pasquale, 2015). Ethical considerations form the bedrock upon which compliance with such regulations can be assured.

1.9. Challenges of Bias and Understanding

Investigating bias in ML-based lending systems is complex, as bias manifests in multifaceted ways:

- **Opacity:** ML models, especially deep learning architectures, often function as "black boxes," making it challenging to pinpoint the exact sources of bias within the decision-making process (Doshi-Velez & Kim, 2017). This lack of explainability hinders bias identification and mitigation.

- **Dynamic Nature:** Bias can stem from historical datasets, feature selection, proxy variables, or even feedback loops created by the model's output. Understanding these multi-layered contributors requires an approach that can disentangle these complexities (Hardt et al., 2016).

- **Intersectionality:** Discrimination often occurs at the intersection of multiple protected attributes like race, gender, and age. This intersectionality demands a nuanced analytical approach to uncover the compounded effects of bias (Crenshaw, 1989).

The challenges and the need for ethical analysis become starkly apparent when considering real-world cases like the Apple Card controversy in 2019, where allegations of gender-based discrimination arose against their credit scoring algorithm (Nellis, 2019).

This research problem statement highlights the inherent ethical dilemmas and the technical complexities involved in understanding and addressing bias in machine learning algorithms used for loan and credit guidance. This dissertation aims to delve

deeply into these challenges, leading to recommendations for ethical and responsible implementations that promote greater fairness in the financial sector.

2. Objectives and Research Questions

2.1. Objectives of the Study

Objective	Description
1	Examine the extent of potential bias in the loan dataset. Quantify disparities in loan approval rates across demographic groups (e.g., gender, marital status, property area).
2	Identify features within the dataset that most strongly correlate with loan approval, investigating if protected attributes (e.g., gender, race, which can be extrapolated from Marital status and/or Dependents) play a disproportionate role.
3	Develop and evaluate bias mitigation techniques to reduce disparities in loan outcomes, specifically focusing on pre-processing and in-processing methods that address the identified biases.
4	Compare the performance of the original ML model with bias-mitigated models, assessing potential trade-offs between fairness metrics and predictive accuracy.

2.2. Research Questions and Hypotheses

Research Question	Hypothesis
Does the ML model trained on the dataset exhibit bias in loan approvals based on protected attributes?	There is a significant statistical difference in loan approval rates between specific demographic groups (e.g., men vs. women, urban vs. rural property areas), even after controlling for other financial factors.
To what extent do specific features in the dataset contribute to biased outcomes?	Features such as gender, marital status, and property area have a disproportionate influence on loan approval decisions, suggesting potential proxy bias.
How effective are different bias mitigation techniques in reducing disparities in loan outcomes?	Bias mitigation strategies (e.g., reweighing data, removing sensitive features, adversarial debiasing) significantly improve fairness metrics (e.g., equalized odds, demographic parity)
What is the impact of bias mitigation techniques on the overall predictive performance of the ML model?	There might be a trade-off between improving fairness metrics and the model's overall accuracy (e.g., slight decrease in accuracy for a significant fairness gain).

3. Literature Review

3.1. Overview

Machine learning (ML) is revolutionising the financial lending industry. Its ability to analyse vast and complex datasets promises efficiency, enhanced risk assessment, and even expanded financial inclusion (Wang, 2021; Lepri et al., 2018). However, using ML in lending also raises critical ethical concerns around fairness, bias, and the need for responsible implementation to avoid discriminatory outcomes.

3.2. Evolution

The use of ML in lending has progressed through distinct stages:

1. **Statistical Risk Scoring:** Early models employed statistical techniques like logistic regression to predict default risk, primarily leveraging structured financial data.

2. **Ensemble Methods and Behavioral Data:** Algorithms like random forests and gradient boosting enhanced predictive performance, and lenders began incorporating behavioural data (e.g., spending habits, social media activity) for alternative credit scoring (The Wall Street Journal, 2015).

3. **Deep Learning and Unstructured Data:** The advent of deep learning empowers models to analyse unstructured data like text and images, potentially expanding access to credit for thin-file borrowers. However, this also raises concerns about explainability and potential new forms of bias.

3.3. Types of Algorithms

Supervised Learning: This is commonly used for loan approvals, where models learn from labelled historical data (e.g., classification algorithms like decision trees, support vector machines, and neural networks).

Unsupervised Learning: Employed for tasks like customer segmentation and anomaly detection in fraud prevention (e.g., clustering algorithms).

Reinforcement Learning: Potential applications in optimising loan pricing and collections strategies through dynamic environmental interactions.

3.4. Advantages

Increased Efficiency and Speed: ML automates many manual processes, leading to faster loan approvals and streamlined operations (Lepri et al., 2018).

Enhanced Risk Assessment: ML can identify complex risk patterns, potentially improving the accuracy of default predictions and optimising credit allocation.

Expanded Financial Inclusion: Alternative data and ML models hold the promise of assessing the creditworthiness of underbanked individuals lacking traditional credit histories (Wang, 2021).

3.5. Challenges

Bias and Fairness: ML algorithms are susceptible to perpetuating existing biases, potentially leading to discriminatory and unjust outcomes, a core focus of this dissertation (Barocas & Selbst, 2016).

Explainability: Complex models (e.g., deep neural networks) can hinder understanding how decisions are made, raising concerns about transparency and accountability (Doshi-Velez & Kim, 2017).

Data Quality and Regulation: Ensuring the quality of data used for training ML models, as well as navigating the evolving regulatory landscape surrounding algorithmic decision-making, are crucial challenges (Citron & Pasquale, 2014).

3.6. Review of Previous Studies and Research

A substantial body of research has investigated the intricate link between ML, bias, and fairness in financial lending. Key areas of investigation include:

- **Identifying and Quantifying Bias:** Numerous studies have empirically demonstrated the presence of bias in lending algorithms. Research indicates that ML models may perpetuate disparities based on race, gender, and other protected attributes, even when those attributes are not explicitly included in the model (Bartlett et al., 2022; Kleinberg et al., 2018).

- **Explaining the Sources of Bias:** Researchers have delved into pinpointing the origins of bias. Studies suggest that historical bias within training data, feature selection processes, and proxy variables can all contribute to discriminatory outcomes (Mehrabi et al., 2021; Hardt et al., 2016).

- **Comparative Analysis of Bias Mitigation Techniques:** Studies have assessed the efficacy of different mitigation strategies. Pre-processing, in-processing, and post-processing methods have all been examined, focusing on their ability to reduce bias while preserving model performance (Hajian et al., 2016; Veale et al., 2018).

- **Ethical and Legal Implications:** Scholars have grappled with the ethical challenges posed by algorithmic bias in lending, exploring the tension between algorithmic decision-making and principles of distributive justice (Citron & Pasquale, 2014). Additionally, there is an ongoing analysis of how existing legal frameworks like the ECOA interact with the complexity of modern ML-based lending (Pasquale, 2015).

3.7. Nuanced Discussion of Fairness Challenges

Bias within ML-based lending systems can manifest in various and interconnected ways:

- **Historical Bias:** ML models trained on data reflecting past discriminatory practices might reproduce and automate these biases, denying credit opportunities to deserving individuals from marginalised communities (Kleinberg et al., 2018).

- **Representation Bias:** If the training data is not fully representative of the population the model will be applied to, it can lead to biased outcomes. For example, a model trained on affluent borrowers might struggle to assess low-income applicants accurately (Hardt et al., 2016).

- **Proxy Bias:** ML models may inadvertently rely on features that act as proxies for protected attributes like race, gender, or age. Using seemingly neutral features like Post Code can perpetuate discrimination if those features are strongly correlated with demographics (Bartlett et al., 2022).

- **Intersectionality:** Bias often intersects across multiple protected characteristics, making it critical to consider how factors like race, gender, and socioeconomic status interact to create compounded disadvantages (Crenshaw, 1989).

3.8. Critical Analysis of Bias Mitigation Techniques

Researchers and practitioners are actively developing strategies to combat bias in ML lending systems:

- **Pre-processing Techniques:** These methods address bias in the data itself before the model is trained. Examples include rebalancing the data to ensure equal representation of groups or removing potentially biased features (Hajian et al., 2016).

- **In-processing Techniques involve** modifying the algorithm's design directly to account for fairness considerations. Approaches include adding fairness constraints to the optimisation objective or employing techniques like

adversarial debiasing (Mehrabi et al., 2021).

- **Post-processing Techniques:** This aims to adjust the model's outputs for greater fairness. Methods include threshold adjustment to alter approval rates across different groups. However, post-processing might limit overall accuracy.

It is vital to note that no single mitigation technique is perfect, and trade-offs often exist between achieving fairness and maintaining predictive accuracy (Kleinberg et al., 2018).

3.9. Regulatory Landscape

Regulatory bodies worldwide are becoming increasingly attentive to the issue of algorithmic bias in lending. Key frameworks and regulations include:

- **Equal Credit Opportunity Act (ECOA) [USA]:** This act prohibits discrimination in lending based on protected attributes. Its applicability to complex ML models remains a subject of active legal debate (Pasquale, 2015).

- **General Data Protection Regulation (GDPR) [EU]:** While emphasising privacy, GDPR provisions like the "right to explanation" have implications for algorithmic transparency in lending (Lepri et al., 2018).

- **Emerging Legislation:** Various bills propose stricter regulations for algorithmic fairness and accountability in financial decision-making.

3.10. Gaps in Knowledge

While extensive, prior research has illuminated several crucial aspects of bias in ML lending systems. However, some gaps create opportunities for further contributions:

- **Limited Real-World Data:** Much of the analysis relies on simulated datasets or publicly available data that may not fully reflect the proprietary models used in real-world lending. This gap highlights the need for studies that can either

directly examine commercial algorithms (if access is possible) or create robust simulations that closely mirror realistic conditions.

- **Focus on Specific Algorithms:** Existing research often focuses on particular algorithm classes (e.g., decision trees, or logistic regression). However, investigations that encompass the broader range of ML techniques used in lending, including deep learning models with their unique challenges of interpretability, are needed.

- **Dynamic Nature of Bias:** Bias must be understood as a dynamic and evolving issue. More research is needed to explore how bias within ML-powered lending systems might change over time due to feedback loops and shifting societal contexts.

3.11. Approaches to Bias Detection

Ensuring the ethical and responsible implementation of artificial intelligence (AI) necessitates actively mitigating bias within machine learning (ML) models deployed for loan and credit guidance systems (Mitchell et al., 2019). A comprehensive literature review requires meticulously examining various strategies to identify these biases.

- **Statistical Disparity Analysis**

Statistical disparity analysis centres on uncovering statistically significant discrepancies in model outcomes (e.g., loan approval rates, interest rates) across protected groups defined by factors such as race, gender, or other demographics (Federal Trade Commission, 2023). This method involves a comparative assessment of metrics between groups. Frequently employed statistical tests include the chi-square test, which assesses the likelihood that observed differences between groups stem from chance (Equation 1), or Fisher's exact test, utilised for smaller sample sizes (Mehta & Patel, 2016).

χ^2 = Σ [(Observed Frequency (O) - Expected Frequency (E))² / Expected Frequency (E)]

where:

χ^2 (Chi-Square): It represents the test statistic to calculate. The larger the chi-square value, the greater the evidence of a statistically significant disparity between the observed and expected frequencies.

Σ (Sigma): We calculate the terms inside the brackets for each category or group in the data, then add them all up to get the overall chi-square statistic.

Observed Frequency (O): This is the actual count of occurrences in each category of the data (e.g., the number of people of a certain race in a specific job category).

Expected Frequency (E): This is the theoretical count expected in each category if there were NO disparity (e.g., if job assignments were completely random with respect to race). The calculation of expected frequencies depends on the specific type of chi-square test are being conducted.

Athey's (2020) study exemplifies this approach. She investigated a loan approval algorithm and identified racial bias using a chi-square test. The test revealed statistically significant discrepancies, demonstrating that the algorithm favoured white applicants over Black applicants with similar creditworthiness.
Careful attention is crucial when refining this approach, particularly in defining protected groups and selecting suitable statistical tests. Power analysis, which determines the probability of detecting an actual effect given a specific sample size and chosen alpha level (typically 0.05), is a valuable tool to ensure sufficient data is available for drawing robust conclusions (Cohen, 1988).

- **Fairness Metrics**

Fairness metrics provide a quantitative measure of the degree of fairness exhibited by model predictions across different groups (Pleiss et al., 2017). Some commonly used metrics include:

Statistical Parity: This metric strives to achieve equal frequencies of positive or negative outcomes (loan approval/denial) across all groups (Equation 2).

SPD = P(Y = 1 | X = A) - P(Y = 1 | X = B)

Where:

P(Y = 1 | X = A) is the probability of a positive outcome (Y = 1) given membership in group A.

P(Y = 1 | X = B) is the probability of a positive outcome (Y = 1) given membership in group B.

Equalized Odds: This metric seeks to equalise the odds of a favourable outcome for all groups, conditioned on their probability of belonging to a protected group (Equation 3).

P(R = + | Y = y, A = a) = P(R = + | Y = y, A = b)

where:

R is the predicted outcome (e.g., positive or negative result)

Y is the true outcome or label (e.g., the actual condition being tested for)

A is the sensitive attribute or group membership (e.g., race, gender, age, etc.)

This equation essentially states that the probability of a positive prediction should be the same for all groups, regardless of their sensitive attribute, given the true outcome. In other words, it means the model should have the same true positive rate (TPR) and false positive rate (FPR) for all groups.

Calibration: This metric assesses how accurately model predictions mirror actual loan default rates across groups (邢 [Xing] et al., 2020). Calibration plots are a standard visualisation tool, where the predicted probability of

default is plotted against the actual observed default rate. Ideally, the data points should fall along a diagonal line, indicating good calibration.

Li et al. (2021) employed a comparative analysis of fairness metrics in their study on bias in a credit scoring model. They calculated statistical parity, equalised odds, and calibration metrics. Their findings revealed that the model exhibited shortcomings in both statistical parity and calibration, suggesting potential bias against certain groups.

Selecting which metrics to utilise is context-dependent on the particular loan or credit guidance scenario. There might be trade-offs between different fairness measures. For instance, achieving perfect statistical parity might lead to unequal odds, and vice versa. Careful consideration is essential to ensure the chosen metric aligns with the specific fairness concept of interest in the loan or credit guidance domain.

- **Feature Importance Analysis**

Feature importance analysis aims to identify features (e.g., zip code, education level) that influence a model's predictions most. This technique can help expose whether features related to protected classes are being given disproportionate weight (Belle & Papantonis, 2021; Rudin et al., 2019). Methodologies such as SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) are instrumental in gauging feature importance. SHAP calculates the contribution of each feature by considering all possible feature combinations, addressing the limitations of simple permutation-based importance techniques. LIME focuses on understanding the model's behaviour at a local level by approximating the model with a simpler, interpretable model around a specific prediction.

Rudin et al. (2019) illustrated the power of this approach when they analysed a student loan repayment prediction model. Their feature importance analysis revealed that zip code, which could be a proxy for socioeconomic status and race, played an excessively influential role in the model's predictions. This highlighted

potential bias, leading to unfair outcomes not directly tied to an individual's creditworthiness.

It is essential to discern the selection of feature importance techniques based on model complexity. It is also worth considering the possibility of complex interactions between features. These interdependencies might mask the true influence of certain variables, necessitating a careful and multifaceted examination.

- **Causal Inference**

Causal inference aims to determine the causal impact of a protected class variable (e.g., race) on the model's results, considering other relevant factors (Kilbertus et al., 2017). This approach requires sophisticated statistical methods such as counterfactual analysis or instrumental variables to arrive at estimations of causal effects.

Obermeyer et al. (2019) demonstrated the utility of causal inference within the context of healthcare risk assessment algorithms. They found that the algorithm assigned lower risk scores to Black patients compared to white patients with the same level of health needs. However, this disparity diminished substantially when controlling for prior healthcare costs, suggesting that the bias primarily stemmed from unequal access to care rather than an inherent racial bias in the algorithm itself.

Developing a strong grasp of research design and domain-specific knowledge is necessary for meticulous fine-tuning and accurate interpretation of causal relationships. Certain data limitations may restrict the applicability of causal inference. For instance, the presence of unobserved confounders can significantly bias causal estimates.

These approaches offer a starting point for detecting bias in loan and credit guidance models. The best approach depends on the specific model, data availability, and desired level of detail.

3.12. Techniques for Bias Detection and Measurement

Effectively addressing bias in machine learning models that guide loan and credit assessments necessitates the employment of a multifaceted suite of detection and measurement techniques. These techniques can be broadly categorised into:

- **Data Analysis Techniques**

Why: Scrutinizing the training data employed to develop an ML model is crucial for pinpointing potential sources of bias. Pre-existing biases or imbalances within the dataset can become entrenched within the model's decision-making (Suresh & Guttag, 2021).

Techniques:

- **Exploratory Data Analysis (EDA):** Involves meticulously examining distributions across various features to uncover group discrepancies. Techniques such as histograms, boxplots, and scatterplots can facilitate visual analysis to unveil differences in income distribution, loan default rates, or other relevant variables across protected demographic groups (Grolemund & Wickham, 2017).

- **Missing Data Analysis:** Uneven patterns of missing data across groups can skew model training and result in biased outcomes. Imputation methods must be carefully selected, as substituting missing values based on group-level averages could worsen existing biases or introduce new ones (Schafer, 1999).

Measurement: Visualize data distributions and missingness patterns for different demographic groups to understand potential biases in the data comprehensively.

Data Reference: The UCI Adult Income Dataset, a commonly used benchmark, contains potential biases related to gender and income. A critical examination of this dataset can illuminate inherent biases that may propagate into models trained upon it (Hajian et al., 2016).

- **Algorithmic Analysis Techniques**

Why: Analyzing the inner workings of the ML model elucidates how it leverages features and arrives at predictions. These insights are essential for assessing the influence of protected class features (which might be directly correlated with protected attributes) within the model's decision-making process.

Techniques:

- **Explainable AI (XAI) Techniques:** Methods such as LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) offer insights into the relative importance of each feature within the model and how they contribute to individual predictions (Lundberg & Lee, 2017; Ribeiro et al., 2016).

Measurement: Employ XAI techniques to generate explanations for model decisions. Through these explanations, you can assess if protected class features, or those strongly correlated with them, exert a disproportionate influence on model outcomes.

Real-Life Example: ProPublica's analysis of the COMPAS algorithm used for criminal risk assessment revealed that the algorithm was significantly biased against Black defendants. XAI techniques could be similarly used to examine whether factors related to race play an excessive role in loan or credit decisions (Angwin, et al., 2016).

- **Counterfactual Analysis**

Why: Counterfactual analysis enables estimation of the model's outcome for a specific individual if they belonged to a different demographic group. This technique seeks to isolate the effect of group membership on a model's predictions, factoring out other contributing variables (Kusner et al., 2017).

Techniques: Counterfactual analysis necessitates the use of advanced causal inference methods and a deep understanding of domain-specific knowledge.

Measurement: Compare the actual model prediction with the counterfactual prediction (generated by altering the protected characteristic) to pinpoint potential bias attributable specifically to group membership.

- **Human-in-the-Loop Analysis**

Why: For high-stakes decisions like loan denials, incorporating human experts helps review model predictions to flag and potentially rectify biased outcomes (Amershi et al., 2014).

Measurement: Systematically log the rate at which human expertise is needed to countermand biased model predictions across different demographic groups. A consistently higher intervention rate for certain groups suggests the presence of bias.

4. Research Methodology

This dissertation employs a systematic literature review (SLR) methodology to critically examine the landscape of bias in machine learning-based loan and credit guidance systems. SLRs offer a structured, transparent, reproducible approach to identifying, evaluating, and synthesising existing knowledge within a particular domain (Kitchenham & Charters, 2007).

4.1. Rational of selecting SLR

An SLR is well-suited for this dissertation's objectives for several reasons:

- **Diverse and Rapidly Evolving Field:** The nexus of machine learning, bias, and fairness in lending is multifaceted and rapidly developing. An SLR will enable a comprehensive synthesis of this expanding body of research.

- **Conceptual and Theoretical Focus:** The primary goal is to map the identified biases, understand the proposed mitigation techniques, and analyse the ethical and regulatory implications. An SLR allows for in-depth theoretical exploration of concepts and debates in the literature.

- **Need for Rigor and Systematicity:** Due to algorithmic bias's ethical and practical significance, a rigorous SLR helps minimise subjective selection biases and promotes greater transparency in the dissertation's findings (Grant & Booth, 2009).

4.2. Alternative methods considered

- **Meta-Analysis:** While valuable for statistically aggregating findings, a meta-analysis may not be feasible, as not all studies on ML bias in lending provide quantifiable effect size measures.

- **Narrative Review:** A traditional narrative review risks less methodological rigour and potential researcher bias in selecting and interpreting literature.

4.3. Pros and Cons of SLR

Pros:

- **Comprehensiveness:** Captures a broad range of studies, potentially reducing bias compared to less structured reviews.

- **Transparency and Replicability:** Explicit search strategy and selection criteria allow for replication and verification of findings.

- **Knowledge Synthesis:** Facilitates identifying patterns, themes, and conflicting viewpoints across the literature.

Cons:

- **Time-Intensive:** The meticulous nature of an SLR can be resource-intensive.

- **Scope Limitations:** While broad, an SLR is ultimately constrained by the available published research.

- **Potential for Publication Bias:** SLRs can be influenced by journals' tendency to publish studies with positive or statistically significant findings.

4.4. Research Questions

RQ1: What types of bias (historical, representational, proxy, etc.) are prevalent in ML-based loan and credit guidance systems?

RQ2: What primary bias mitigation strategies are proposed in the literature, and how is their effectiveness evaluated?

RQ3: How do ethical frameworks address the issue of bias in ML lending, and what are the regulatory debates surrounding algorithm fairness?

4.5. Scope, Inclusion & Exclusion Criteria

Criteria	Inclusion	Exclusion
Time Period	Studies published from 2018 to the present.	Studies prior to 2018
Document Type	Reviewed journal articles, conference papers	White papers, blog posts, non-academic sources
Focus	ML bias in lending, fairness, regulation	ML in other domains, without clear lending focus
Language	English	Other Languages

4.6. Study Selection

- **Academic Databases:** Google Scholar, Web of Science, IEEE Xplore, ACM Digital Library

- **Relevant Journals:** ACM Computing Surveys, Machine Learning, Proceedings of the Conference on Fairness, Accountability, and Transparency (ACM FAT*)

4.7. Search Sequences

Database	Search String
Google Scholar (https://scholar.google.com/)	("machine learning" OR "AI") AND ("lending" OR "credit") AND ("bias" OR "fairness" OR "discrimination" OR "ethics")
ACM Digital Library (https://dl.acm.org/)	("machine learning" OR "deep learning") AND "loan" AND ("fair" OR "bias")
IEEE Xplore (https://ieeexplore.ieee.org/)	("machine learning" OR "artificial intelligence") AND "credit decision" AND ("algorithmic fairness" OR "bias mitigation")

Search Refinement Strategies

- **Boolean Operators:** Use 'AND', 'OR', and 'NOT' to combine and refine search terms.

- **Truncation/Wildcards:** Use “" to search for variations of a word stem (e.g., "discrim" for discrimination, discriminatory, etc.).

- **Synonyms and Related Terms:** Expand your keywords to capture different ways bias and fairness might be described.

- **Citation Tracking** Use Google Scholar's "Cited by" feature to track research citing seminal papers identified in your initial search.

- **Snowballing:** Examine bibliographies of relevant articles to identify other potentially valuable studies.

4.8. Prioritisation Table

Author(s)	Year	Key Findings	Relevance Score (1-5)
Mehrabi et al.	2021	Comprehensive survey of bias types present in ML systems, bias mitigation techniques, and their evaluation metrics.	5 (Highly Relevant)
Bartlett et al.	2022	Investigates discrimination in FinTech lending, specifically focusing on proxy bias and disparate outcomes.	4 (Relevant)
Kleinberg et al.	2018	Theoretical exploration of the inherent trade-offs between fairness metrics and predictive accuracy in risk scoring.	3 (Moderately Relevant)
Hajian et al.	2016	Early investigation of pre-processing techniques for bias mitigation in data.	3 (Moderately Relevant)

How to Build This Table

1. **Gather Search Results:** Compile the top relevant hits from each database search.

2. **Summarise Key Findings:** Concisely capture each paper's primary contributions and focus in 1-2 sentences.

3. **Assign Relevance Score:** Rate each paper on its direct relevance to your research questions (a simple 1-5 scale works well).

4. **Notes:** Add brief comments on why the work is essential, any limitations, or connections to other papers.

Prioritisation Strategy

Highest Relevance First: Focus on the papers with scores of 4 and 5.

Balance: Include a mix of theoretical, empirical, and review papers for a well-rounded understanding.

Iterative: Update this table as your searches expand and your understanding of the field deepens.

Caveats

Subjective Scores: Assign relevance scores meaningfully to your specific dissertation focus.

Search is Ongoing: This is a snapshot; continuously add and adjust as you expand your SLR.

4.9. Threats to Validity

Researcher Bias: Reflexivity and using a co-reviewer will help mitigate potential biases in study selection.

Publication Bias: Efforts will be made to identify relevant gray literature (e.g., reports)

Evolving Terminology: The search strategy will be designed to capture variations in terminology used to describe bias and fairness.

5. Scope and Limitations

5.1. Scope of the Study

Dataset: The primary analysis will utilise the "Training Loan Data" dataset from Kaggle (https://www.kaggle.com/datasets/jboysen/ny-home-mortgage/download/ny_hmda_2015.csv). While this allows for robust exploration of bias and mitigation, it is essential to acknowledge the potential limitations of a single dataset.

Protected Attributes: The study will investigate potential biases related to gender, marital status, and property area. It is acknowledged that other sensitive attributes (e.g., race, age) might not be directly available in this dataset but could be inferred indirectly.

Algorithm Focus: The main emphasis will be on supervised machine learning algorithms for loan approval classification (e.g., decision trees, logistic regression, potential neural networks). While some insights might be transferable, the study will not explicitly delve into other algorithmic areas, such as unsupervised or reinforcement learning.

Mitigation Strategies: The study will investigate bias mitigation techniques in pre-processing (e.g., data rebalancing) and in-processing (e.g., adversarial debiasing, fairness constraints). Post-processing methods might be discussed, but not the core experimental focus.

5.2. Limitations and Constraints

Dataset Representativeness: The Kaggle dataset may have inherent biases or not fully reflect real-world lending scenarios. Findings will need to be interpreted considering these dataset-specific limitations.

Interpretability vs. Performance: Certain ML models (e.g., deep neural networks) offer excellent predictive power but can lack transparency. The study may encounter trade-offs between understanding the mechanics of bias and maximising predictive accuracy.

Ground Truth: Without access to the proprietary models used by financial institutions, perfectly replicating real-world bias scenarios is difficult. The research will use simulations and well-defined bias metrics as proxies.

Evolving Regulations: The legal and regulatory landscape around algorithmic bias is in flux. While the study will refer to existing frameworks (e.g., ECOA), it may not fully anticipate future regulatory developments.

6. Evaluation Metrics and Analysis Techniques

Machine learning (ML) models can inadvertently perpetuate harmful biases in training data, leading to discriminatory and unfair outcomes. Robust evaluation metrics and analysis are paramount to understanding and counteracting these biases.

6.1. Evaluation Metrics

Bias Detection Metrics

These quantify disparities between groups (often protected classes) in the model's outputs.

For example, "Equalized odds aims to ensure that individuals from different protected groups have similar probabilities of being correctly classified, regardless of whether the true outcome is positive (loan approval) or negative (loan denial). This balances both false positives (incorrectly predicted to repay) and false negatives (incorrectly predicted to default), promoting a fairer distribution of both benefits and burdens."

Statistical Parity/Demographic Parity: Compares the proportion of favourable outcomes (loan approvals) across groups. A significant difference suggests potential bias (Hajian et al., 2016).

> Example: If women have a 60% loan approval rate and men have an 80% rate, this signals a disparity.

Equalised Odds: Measures whether protected groups have equal rates of true positives and false positives. This addresses situations where overall accuracy might be high but hides discrimination (Hardt et al., 2016).

> Example: Both men and women should have similar rates of being correctly classified as likely to repay and incorrectly classified as likely to default.

Disparate Impact: Focuses on the ratio of favourable outcomes between groups. Legally relevant in contexts like the ECOA (Pasquale, 2015). They are often expressed as the "80% rule" (a minority group must receive favourable outcomes at least 80% as often as the majority).

Equal Opportunity: Focuses specifically on true positive rates—the chance of a qualified individual from a protected group receiving a positive outcome. Disparities in equal opportunity suggest bias (Hardt et al., 2016).

6.2. Mitigation Assessment Metrics

These gauge how well a mitigation technique reduces bias while considering other performance factors:

- **Change in Bias Metric:** Directly measure if the chosen bias metric (e.g., equalised odds) improves after mitigation.

- **Accuracy Trade-offs:** Compare the model's overall accuracy (or other performance metrics like precision/recall) before and after mitigation. This is crucial, as some fairness interventions might harm predictive capabilities.

- **Group-specific Accuracy:** Checks if accuracy for any protected group declines significantly post-mitigation. Sometimes, improving overall fairness can unintentionally worsen performance for specific subgroups.

6.3. Bias Mitigation Techniques

Pre-processing:

- **Data rebalancing:** Oversampling underrepresented groups or undersampling overrepresented groups to achieve a more balanced dataset (Kamiran & Calders, 2012).

- **Reweighing:** Assigning weights to data instances to reduce the influence of overrepresented groups (Jiang & Nachum, 2020).

In-processing:

- **Adversarial debiasing:** Introducing an adversarial component that learns to predict the protected attribute while the primary model tries to perform its task

without using the protected attribute's information (Zhang et al., 2018).

- **Regularisation:** Applying penalties to model parameters associated with discriminatory patterns (Kamishima et al., 2012).

Statistical Tests: To quantify disparities statistically, employ hypothesis testing (e.g., Chi-squared tests) or regression analysis with protected attributes as variables.

Visualisations: Boxplots, histograms, and scatter plots effectively visualise distributions of outcomes and potential biases across groups.

Explainability Techniques: Tools like SHAP (Shapley Additive Explanations) help understand how specific features, including potentially biased ones, contribute to model predictions on an individual level (Lundberg & Lee, 2017).

6.4. Bias Assessment

Fairness Audits: Comprehensive examinations of ML systems across their lifecycles to identify potential sources of bias. These audits may include technical testing and qualitative assessments of the system's social and ethical implications (Selbst et al., 2019).

Counterfactual Explanations: Explore "what-if" scenarios. By changing features like a protected attribute, one can observe how it affects the model's prediction, revealing potential biases (Wachter et al., 2017).

6.5. Important Considerations

Context Matters: The appropriate choice of metrics and mitigation techniques depends heavily on the ML system's specific application domain and social implications.

There is no Single Solution: Bias mitigation remains an active research area. Often, a combination of strategies is crucial.

Tradeoffs: Bias mitigation can sometimes slightly impact model performance. It is essential to find a balance between fairness and overall accuracy.

6.6. Ethical Considerations in Bias Mitigation

The presence of bias in machine learning systems raises profound ethical concerns that go beyond technical solutions. Here is a breakdown of critical ethical considerations and their significance:

Fairness and Non-Discrimination:

- **Meaning:** Bias mitigation upholds the principles of fairness and non-discrimination by ensuring that protected attributes like race, gender, ethnicity, religion, etc., do not lead to unjust or unequal treatment for individuals or groups (Mehrabi et al., 2021).

- **Addressing Bias:** Bias mitigation techniques aim to create models that make decisions without considering protected attributes or reducing the influence of historical biases encoded in the data.

- **Importance:** Fairness is essential for preserving fundamental human rights and preventing the perpetuation of societal inequities. Biased AI systems can deny marginalised groups opportunities, resources, or services, exacerbating disparities.

Justice and Social Responsibility:

- **Meaning:** Developers and deployers of ML models hold a social responsibility to ensure their systems do not cause harm or perpetuate social injustice. This involves recognising potential bias and proactively implementing mitigation strategies (Selbst et al., 2019).

- **Addressing Bias:** Bias mitigation necessitates carefully considering how models are designed and used. It means being intentional about measuring success and the potential downstream consequences, especially for

historically disadvantaged groups.

- **Importance:** Without a commitment to justice, biased AI systems can reinforce existing power structures, leading to further discrimination and exclusion.

Transparency and Explainability:

- **Meaning:** Transparency involves being open about data collection, model development, and the rationale behind models' decisions. Explainability means understanding how the model arrived at its outputs (Barocas et al., 2019).

- **Addressing Bias:** Transparency and explainability allow stakeholders to scrutinise models, identify potential sources of bias, and hold developers accountable.

- **Importance:** Without transparency, assessing whether bias exists and taking corrective actions is difficult. Explainable models foster trust and enable challenges to decisions based on discriminatory patterns.

Accountability and Governance:

- **Meaning:** There must be precise mechanisms for accountability in developing and deploying ML systems. This includes establishing governance structures that identify responsible parties and enable redress when biases cause harm (Rahwan, 2018).

- **Addressing Bias:** Effective governance and accountability frameworks incentivise developers and institutions to prioritise bias mitigation and ensure that there are consequences for deploying biased systems.

- **Importance:** Accountability is crucial for building public trust in AI and ensuring those who cause harm through biased systems are held responsible.

Human Oversight and Values:

- **Meaning:** AI systems should never fully replace human judgment, especially in high-stakes domains. Human values, ethics, and contextual understanding must always guide the use of such technology (Leslie, 2019).

- **Addressing Bias:** Human oversight helps contextualise model outputs, catch unintended biases, and intervene when decisions conflict with ethical principles.

- **Importance:** Humans are essential in identifying subtle biases that algorithms may miss and ensuring that AI systems align with societal values.

Bias mitigation is not simply a technical challenge—it is fundamentally intertwined with ethical considerations. Addressing these considerations is critical for ensuring AI systems benefit society fairly and justly.

7. Implementation and Experimentation

7.1. Data Collection and Preprocessing

The data collection and pre-processing will be done as per the following steps as outlined below:

Step	Description	Program details
Data Collection	Obtains the raw data essential for your analysis and model building.	Loading the CSV file using pandas.read_csv().
Exploratory Analysis	Provides an initial understanding of the data's structure, quality, and potential issues.	- df.head() to examine initial rows. - df.info() to check data types and missing values. - df.describe() to see statistical summaries.
Data Cleaning	Addresses inconsistencies, errors, and missing data to improve data quality.	- Handling missing values (e.g., fillna, deletion). - Correcting inconsistent entries.
Feature Encoding	Converts categorical features into numerical representations suitable for ML algorithms.	- Selecting categorical columns. - Applying encoders (e.g., OneHotEncoder, LabelEncoder)
Feature Scaling	Standardizes or normalizes the range of numerical features to prevent bias in some models.	- Selecting numerical columns. - Applying scalers (e.g., StandardScaler, MinMaxScaler).
Target Variable Preparation	Ensures the target variable is in a format compatible with ML algorithms.	- Encoding if necessary (e.g., converting 'Y'/'N' to 0/1).
Identifying Protected Attributes	Explicitly defines which features relate to protected groups for bias analysis.	- Selecting columns based on your research focus and ethical considerations.
Data Splitting	Divides the dataset into training and testing sets for model building and evaluation.	- Using train_test_split() - Specifying test size and random state for reproducibility.

7.2. Model Development and Training

The below steps will be followed for Model Development and Training:

Step	Description	
Model Selection	Chooses an appropriate ML algorithm suitable for classification and bias analysis.	- Researching different algorithms (e.g., Logistic Regression, Decision Trees, Random Forests). - Considering factors like interpretability, desired bias metrics.
Model Training	Fits the model to the training data to learn patterns between features and loan outcomes.	- Using the fit() method of the chosen model on X_train and y_train.
Model Evaluation (Baseline)	Assesses performance on unseen data to gauge accuracy and potential overfitting.	- Using predict() to generate predictions on X_test. - Calculating performance metrics (accuracy, precision, recall, F1-score).
Bias Detection	Employs statistical measures to quantify disparities in model outcomes across protected groups.	- Calculating metrics like statistical parity, equalized odds, disparate impact. - Comparing these metrics across groups defined by protected attributes.
Feature Importance Analysis	Identifies the features that have the most significant influence on the model's predictions.	- Techniques like permutation importance or SHAP values.

7.3. Implementation and Analysis

The implementation and analysis are used using the dataset from Kaggle and then doing analytical exercises using R (R-Studio) and Python.

7.4. Artefacts

The summarised analysis was used using Jupyter and Python in actuals and it can be found at https://github.com/biswassandip/dissertation/blob/main/src/data-analysis/eda_assessment.ipynb

To achieve the results further analysis was done by experimenting with the dataset using R (R-Studio) and running complex calculations to simplify trees like CART and getting a simplified result for explanations.

Python version: 3.10.2

Libraries: The libraries used are provided below and the highlighted ones were used for most of the analysis. All libraries used can be found here, https://github.com/biswassandip/dissertation/blob/main/src/data-analysis/requirements.txt

Library Name	Description	Reference URL
fairlearn	A library for assessing and mitigating unfairness in machine learning models.	https://pypi.org/project/fairlearn/
graphviz	A library for creating and rendering graphs.	https://pypi.org/project/graphviz/
matplotlib	A comprehensive library for creating static, animated, and interactive visualizations in Python.	https://pypi.org/project/matplotlib/
matplotlib-inline	A library for enabling inline plotting in Jupyter notebooks.	https://pypi.org/project/matplotlib-inline/
mizani	A library for providing scales and aesthetic mappings for the plotnine plotting library.	https://pypi.org/project/mizani/
numpy	A fundamental package for scientific computing with Python.	https://pypi.org/project/numpy/
pandas	A powerful library for data analysis and manipulation.	https://pypi.org/project/pandas/
pandocfilters	A library for creating pandoc filters in Python.	https://pypi.org/project/pandocfilters/
patsy	A library for describing statistical models (especially linear models) and building design matrices.	https://pypi.org/project/patsy/
Pillow	A library for image processing.	https://pypi.org/project/pillow/

Plotly	A library for creating interactive visualizations.	https://pypi.org/project/plotly/
plotnine	A grammar of graphics for Python based on ggplot2.	https://pypi.org/project/plotnine/
scikit-learn	A machine learning library featuring various classification, regression, and clustering algorithms including support vector machines, random forests, gradient boosting, k-means, and DBSCAN.	https://scikit-learn.org/stable/
seaborn	A data visualization library based on matplotlib that provides a high-level interface for drawing attractive and informative statistical graphics.	https://seaborn.pydata.org/
Shap	A library for explaining the output of any machine learning model. SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model.	https://shap.readthedocs.io/en/latest/
xgboost	An optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework.	https://xgboost.readthedocs.io/en/stable/

Dataset: https://www.kaggle.com/datasets/jboysen/ny-home-mortgage/download/ny_hmda_2015.csv

7.5. Dataset

The dataset can be found here, https://www.kaggle.com/datasets/jboysen/ny-home-mortgage/download/ny_hmda_2015.csv.

The references about the dataset can be found here, https://www.consumerfinance.gov/data-research/hmda/learn-more.

The Home Mortgage Disclosure Act (HMDA), enacted by the United States Congress in 1975, mandates that a wide array of financial institutions collect, report, and

publicly disclose granular data regarding their mortgage lending activities (Consumer Financial Protection Bureau, n.d.). This comprehensive dataset serves multiple critical functions.

Firstly, HMDA data provides valuable insights into whether lenders are adequately meeting the housing credit needs of their local communities (Charles & Hurst, 2002). By analyzing lending patterns across different geographic areas and demographic groups, researchers and policymakers can identify potential disparities in access to mortgage credit (Fishbein, 1981).

Secondly, HMDA data equips public officials with essential information to inform their decision-making processes and shape housing policies (Federal Financial Institutions Examination Council, n.d.). For instance, by examining the distribution of loan approvals and denials, policymakers can assess the effectiveness of fair lending laws and regulations (Charles & Hurst, 2002).

Lastly, HMDA data plays a pivotal role in identifying discriminatory lending practices (Canner et al., 1995). By scrutinizing loan terms and conditions across different borrower characteristics, researchers can detect potential instances of redlining or other forms of unlawful discrimination (Ladd, 1982).

However, it is important to note that the publicly available HMDA data is carefully modified to safeguard the privacy of applicants and borrowers (Consumer Financial Protection Bureau, 2018). This ensures that individuals' personal and financial information remains confidential while still allowing for meaningful analysis of aggregate lending patterns (Federal Financial Institutions Examination Council, 2023).

7.6. Terminologies and data perspectives

Before starting with the analysis and techniques, let us understand some of the data perspectives and terminologies used within the implementation

- **lien status** refers to the legal claim a lender has on a borrower's property. This claim serves as collateral for the loan and gives the lender the right to seize and sell the property if the borrower fails to repay the loan as agreed.

- **Loan originated** refers to when a loan is closed. The last stage of the loan is Loan Origination.

- **Location** encompasses the geographical identifiers of the subject property, including the state, metropolitan statistical area (MSA), and census tract designation.

- **Property Type** in the context of mortgage lending data encompasses two distinct aspects: the physical structure of the property and its intended occupancy. This categorical variable typically includes three main values for property type:

    - **One-to-four family dwelling:** This category refers to residential properties designed to accommodate one to four families, such as single-family homes, duplexes, triplexes, and quadruplexes.
    - **Manufactured housing:** This category encompasses prefabricated homes, including mobile homes and modular homes, which are constructed in a factory and then transported to their final location.
    - **Multifamily dwelling:** This category includes residential properties designed for multiple families, such as apartment buildings and condominiums with five or more units.

- **Applicant**: The HMDA data includes demographic information for both applicants and co-applicants on mortgage loan applications. This encompasses the gender and racial/ethnic background of each individual. Specifically, the data identifies the sex of the applicant and co-applicant, as well as the racial and ethnic identity of each. This information is collected to monitor fair lending practices and identify potential disparities in access to mortgage credit.

7.7. Dataset

The dataset comprises of 439653 entries and 78 columns. Refer github to execute and refer to all the columns within the dataset.

Some of the columns that have been taken into consideration are provided with relevant info.

lien_status		
1	lien_status_1	Secured by a first lien
2	lien_status_2	Secured by a subordinate lien
3	lien_status_3	Not secured by a lien
4	lien_status_4	Not applicable

action_taken		
1	action_taken_1	Loan originated
2	action_taken_2	Application approved but not accepted
3	action_taken_3	Application denied by financial institution
4	action_taken_4	Application withdrawn by applicant
5	action_taken_5	File closed for incompleteness
6	action_taken_6	Loan purchased by the institution
7	action_taken_7	Preapproval request denied by financial institution

property_type		
1	property_type_1	One-to-four family dwelling (other than manufactured housing)
2	property_type_2	Manufactured housing
3	property_type_3	Multifamily dwelling

applicant_ethnicity		
1	applicant_ethnicity_1	Hispanic or Latino
2	applicant_ethnicity_2	Not Hispanic or Latino
3	applicant_ethnicity_3	Information not provided by applicant in mail, Internet, or telephone application
4	applicant_ethnicity_4	Not applicable

applicant _race_1		
1	applicant _race_1_1	American Indian or Alaska Native
2	applicant _race_1_2	Asian
3	applicant _race_1_3	Black or African American
4	applicant _race_1_4	Native Hawaiian or Other Pacific Islander
5	applicant _race_1_5	White
6	applicant _race_1_6	Information not provided by applicant in mail, Internet, or telephone application

7	applicant _race_1_7	Not applicable

co_applicant_ethnicity		
1	co_applicant_ethnicity_1	Hispanic or Latino
2	co_applicant_ethnicity_2	Not Hispanic or Latino
3	co_applicant_ethnicity_3	Information not provided by applicant in mail, Internet, or telephone application
4	co_applicant_ethnicity_4	Not applicable
5	co_applicant_ethnicity_5	No co-applicant

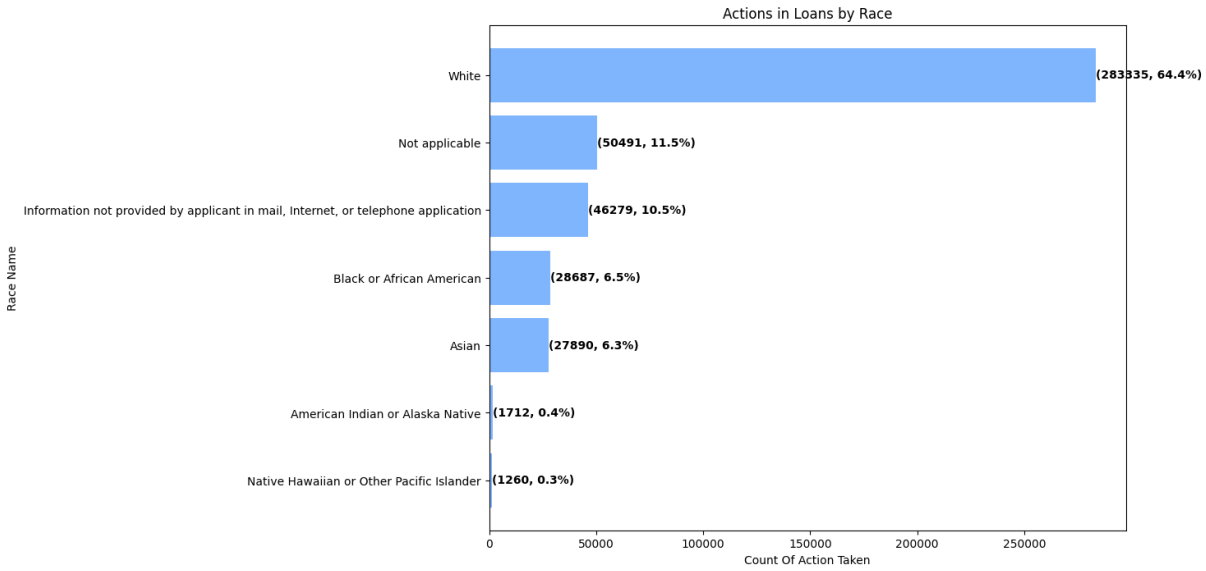
co_applicant_race_1		
1	co_applicant_race_1_1	American Indian or Alaska Native
2	co_applicant_race_1_2	Asian
3	co_applicant_race_1_3	Black or African American
4	co_applicant_race_1_4	Native Hawaiian or Other Pacific Islander
5	co_applicant_race_1_5	White
6	co_applicant_race_1_6	Information not provided by applicant in mail, Internet, or telephone application
7	co_applicant_race_1_7	Not applicable
8	co_applicant_race_1_8	No co-applicant

7.8. Data Analysis

Actions on Loans distribution

Analyse the distribution of various loan actions within the dataset. As previously established, our primary interest lies in the "Loan Origination" action, as it indicates that the loan application has been approved and the funds are being disbursed to the applicant. A comprehensive understanding of the prevalence of this action is crucial for assessing the overall lending activity and its impact on borrowers and communities.

Inference: *It is observed that more than 50% of the loans are originated.*



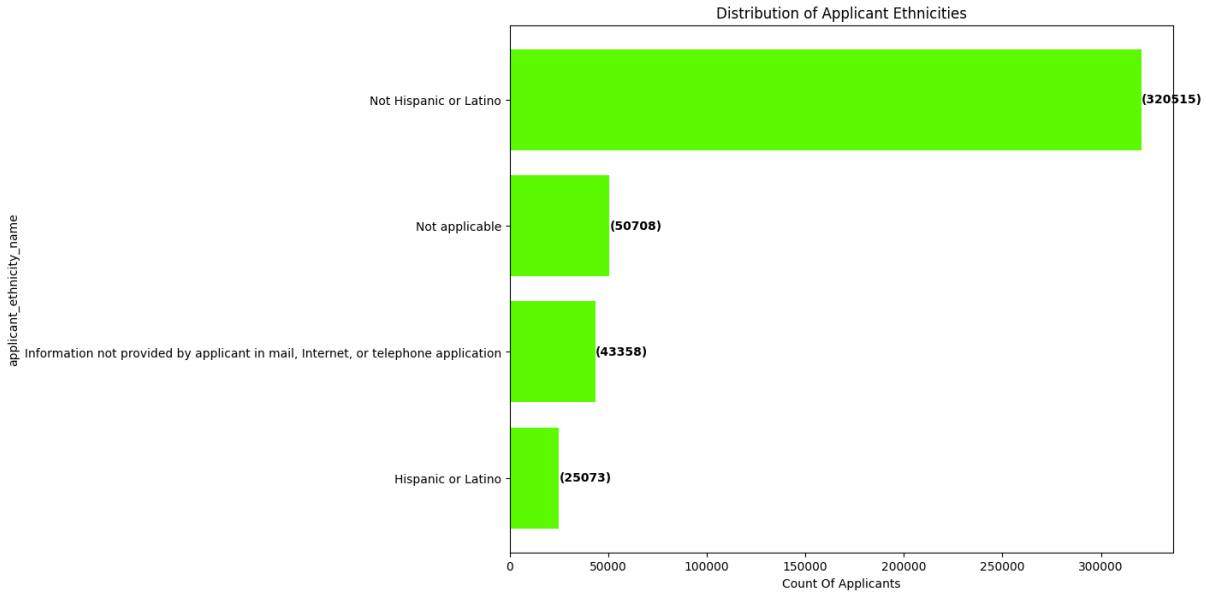
Dependencies of Ethnicity on the Actions

Evaluation of various ethnic groups that have been found associated with loan origination.

Distribution of Ethnic Groups

The graph below shows the distribution of various ethnic groups.

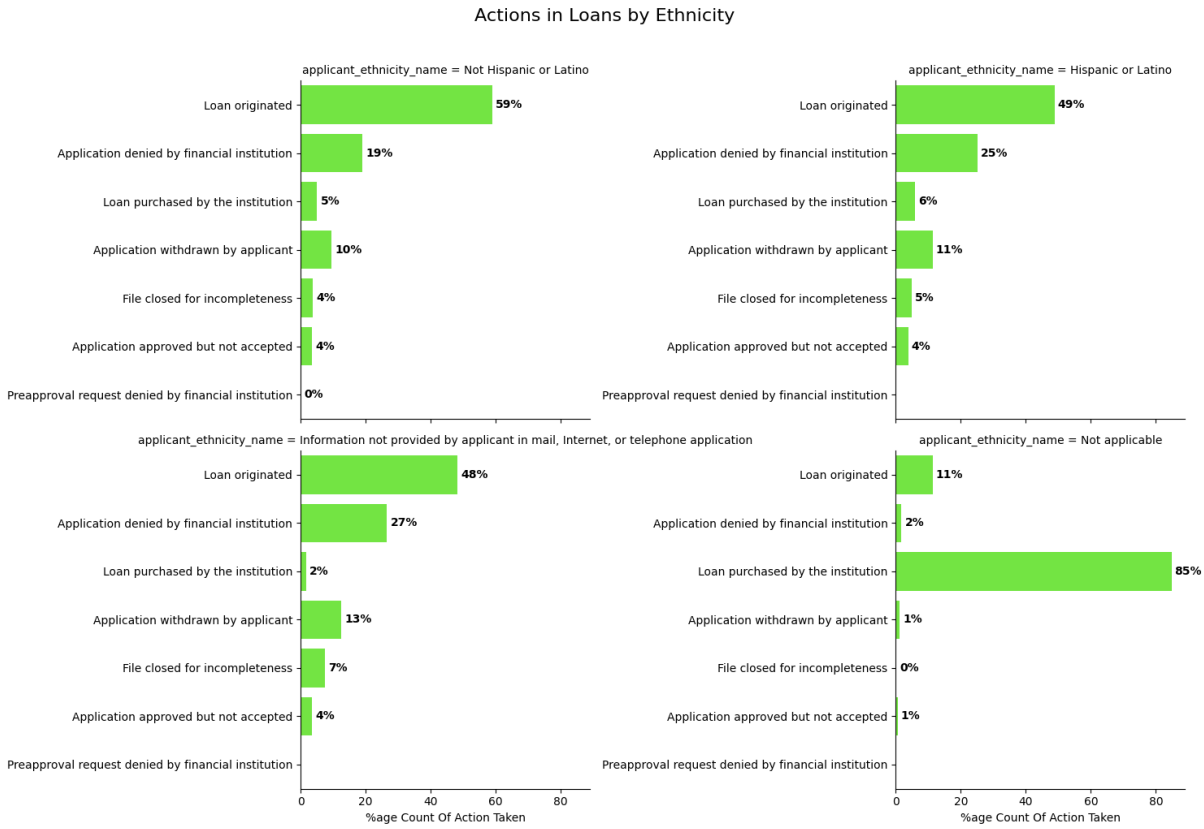
Inference: *The "Not Hispanic or Latino" comprises to the largest percentage of the loans.*



Loan Status vs Ethnicity

This distribution shows the impact on loan status per ethnicity.

Inference: *The analysis reveals a notable disparity in loan outcomes between the "Not Hispanic or Latino" and "Hispanic or Latino" ethnic communities. The "Not Hispanic or Latino" community exhibits a higher proportion of loan originations (59%) compared to the "Hispanic or Latino" community (49%). Conversely, the "Not Hispanic or Latino" community experiences a lower rate of applications denied by financial institutions (19%) compared to the "Hispanic or Latino" community (25%). This discrepancy suggests potential disparities in access to credit and differential treatment by lenders based on ethnicity.*

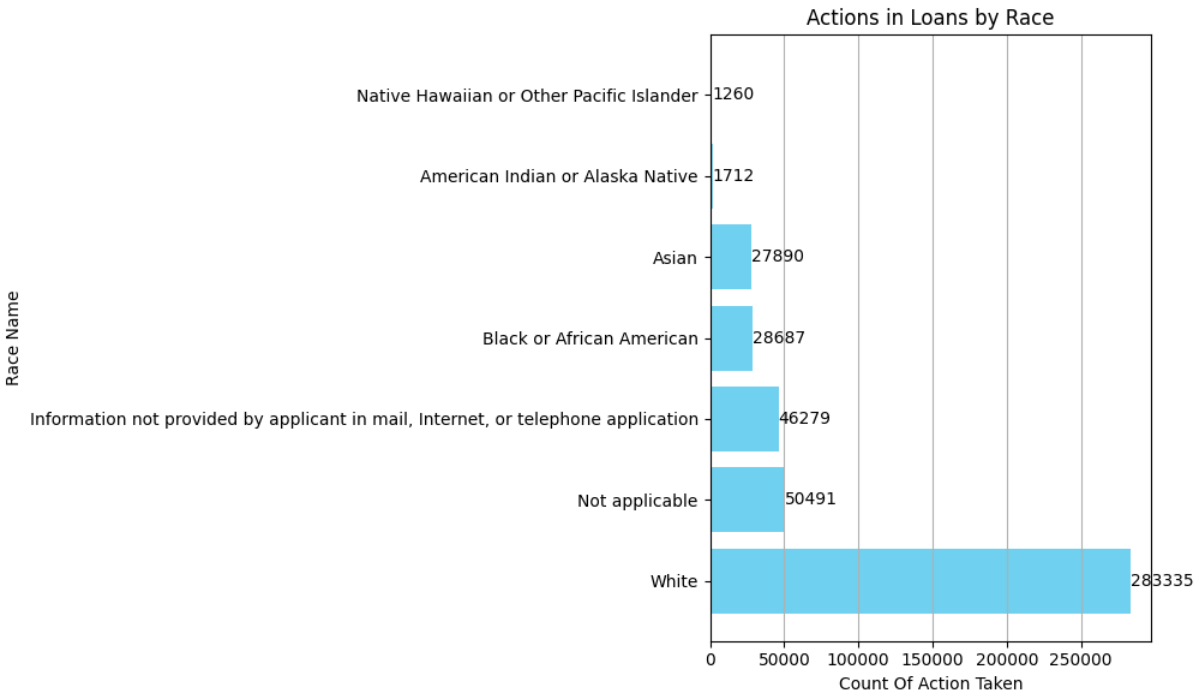


Loan origination impacts to Race

Races

The below distribution explores the various races that are related to the loan origination process.

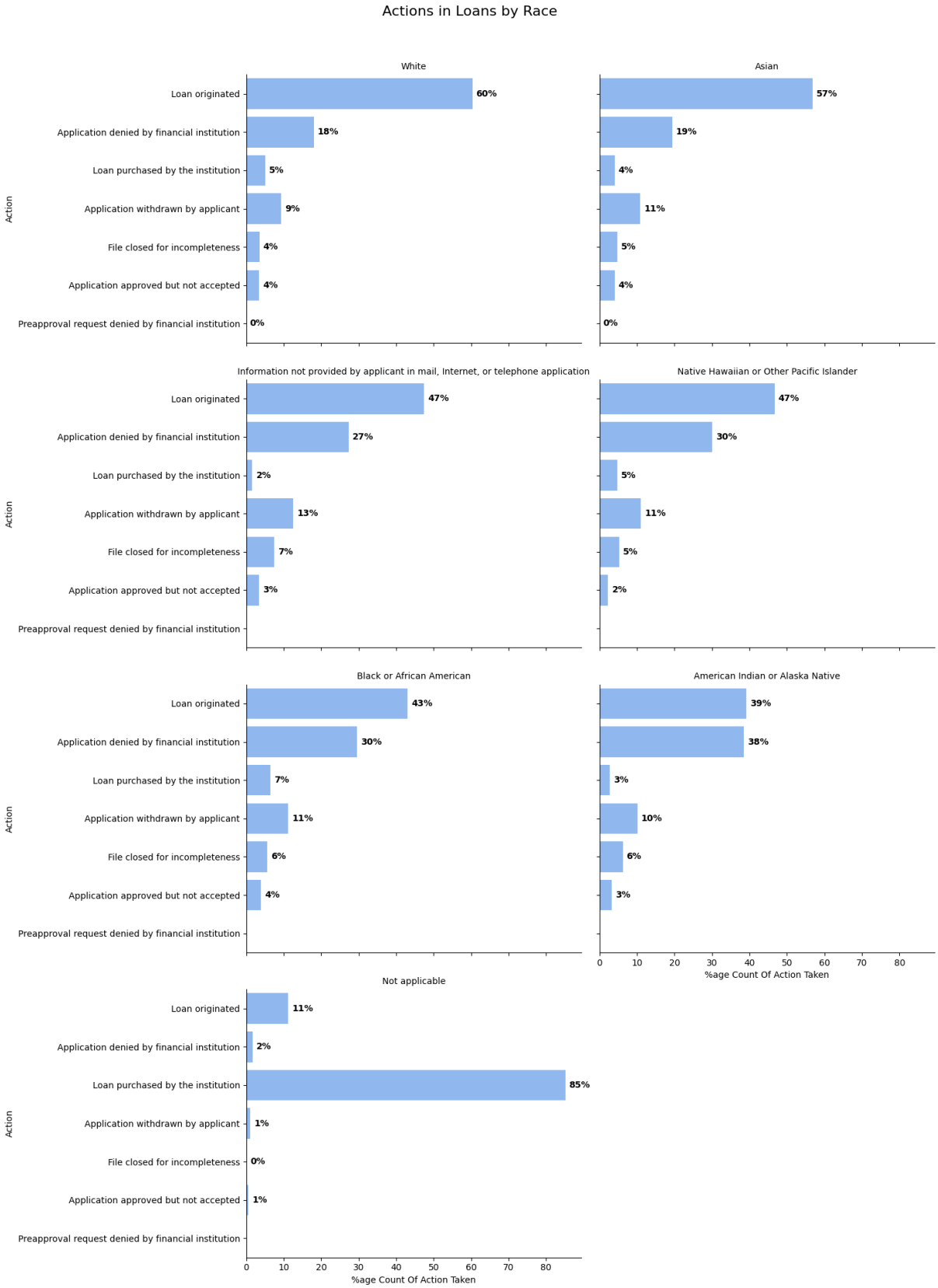
Inference: *It is observed that the white community is associated with the largest percentage of loans.*



Loan Status vs Race

The graphs below provide information about the action that is associated with each race.

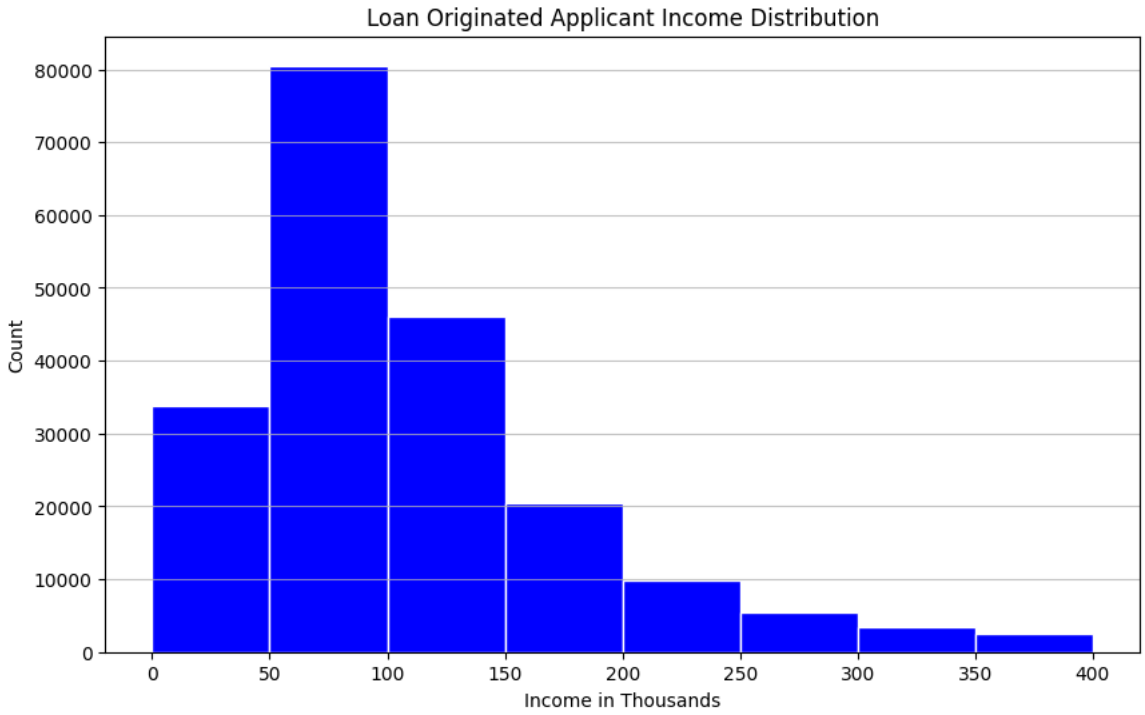
Inference: *The analysis reveals a higher percentage of loan originations within the White and Asian communities compared to the Black or African American community. This discrepancy suggests potential disparities in access to mortgage credit based on racial background and warrants further investigation into the underlying factors contributing to this observed difference.*


Actions in Loans by Race
White
Asian
Information not provided by applicant in mail, Internet, or telephone application
Native Hawaiian or Other Pacific Islander
Black or African American
American Indian or Alaska Native
Not applicable
Action
Loan originated
Application denied by financial institution
Loan purchased by the institution
Application withdrawn by applicant
File closed for incompleteness
Application approved but not accepted
Preapproval request denied by financial institution
60%
18%
5%
9%
4%
4%
0%
57%
19%
4%
11%
5%
4%
0%
47%
27%
2%
13%
7%
3%
47%
30%
5%
11%
5%
2%
43%
30%
7%
11%
6%
4%
39%
38%
3%
10%
6%
3%
11%
2%
85%
1%
0%
1%
0
10
20
30
40
50
60
70
80
%age Count Of Action Taken

Applicant vs Income Distribution

This is to find out the income limits and its distribution for the applicants.

Inference: *The analysis reveals a significant concentration of loan originations among applicants with incomes ranging from $60,000 to $80,000. This suggests that borrowers within this income bracket represent a substantial portion of successful mortgage applicants. Further investigation is warranted to understand the factors contributing to this pattern, including the affordability of housing options and lending practices within this income range.*



Loan Purpose Types

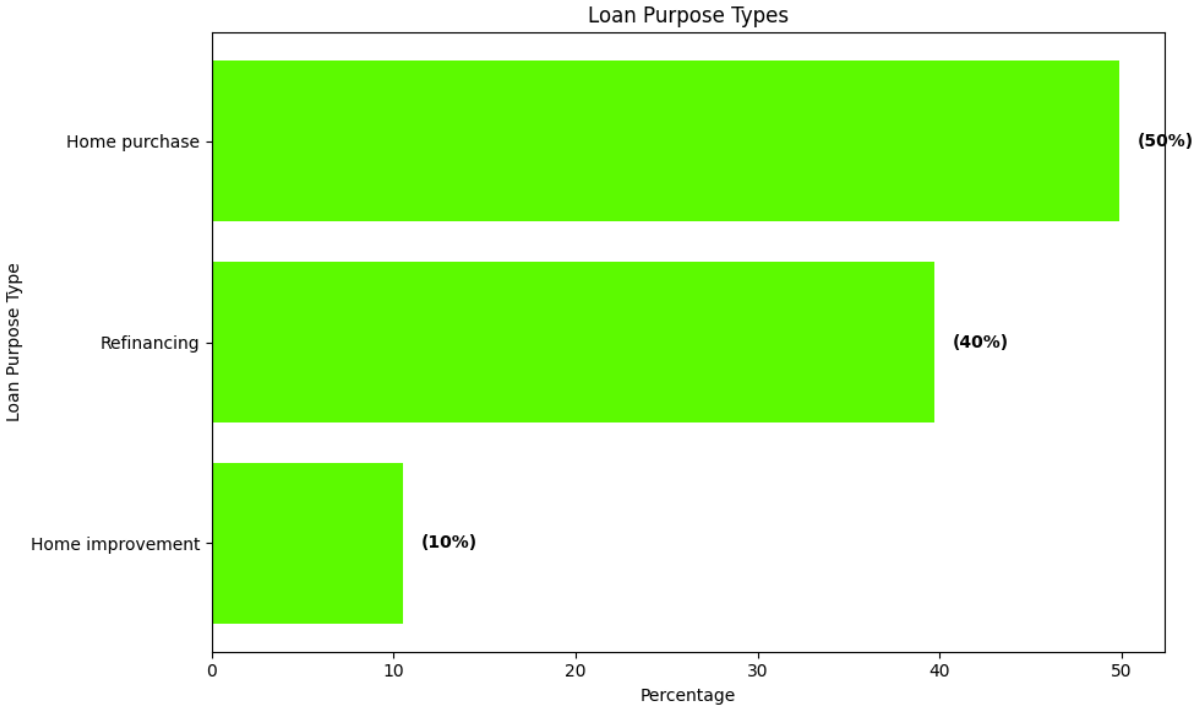
The analysis proceeds with an investigation into the various loan purpose types associated with the loans in the dataset.

An examination of the distribution and characteristics of these loan purposes will shed light on the underlying reasons for borrowing and potentially reveal any disparities or trends across different demographic groups.

Distribution of Loan Purpose Types

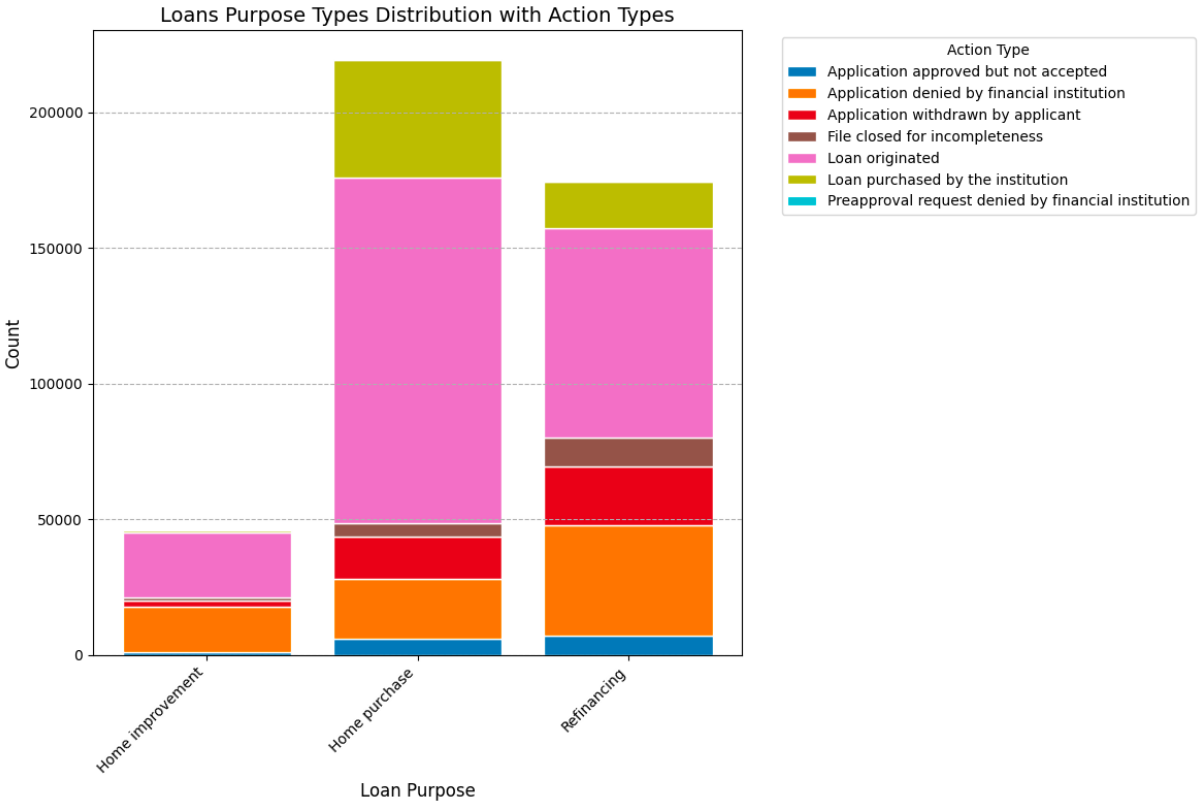
The graph below depicts the distribution of loan purpose types.

Inference: *The predominant loan purpose categories within the dataset are "Home Purchase" and "Refinancing." These categories represent the primary reasons borrowers seek mortgage loans, either to finance the acquisition of a new property or to modify the terms of an existing mortgage, respectively.*



Loan purpose types vs actions

The following bar graph presents a visual representation of the relationship between loan purpose types and the corresponding loan actions. It provides an overview of the distribution of different loan purposes across various stages of the lending process, including origination, denial, and other outcomes. This graphical representation facilitates a comprehensive understanding of the interplay between loan purpose and loan action, offering insights into potential trends and disparities.

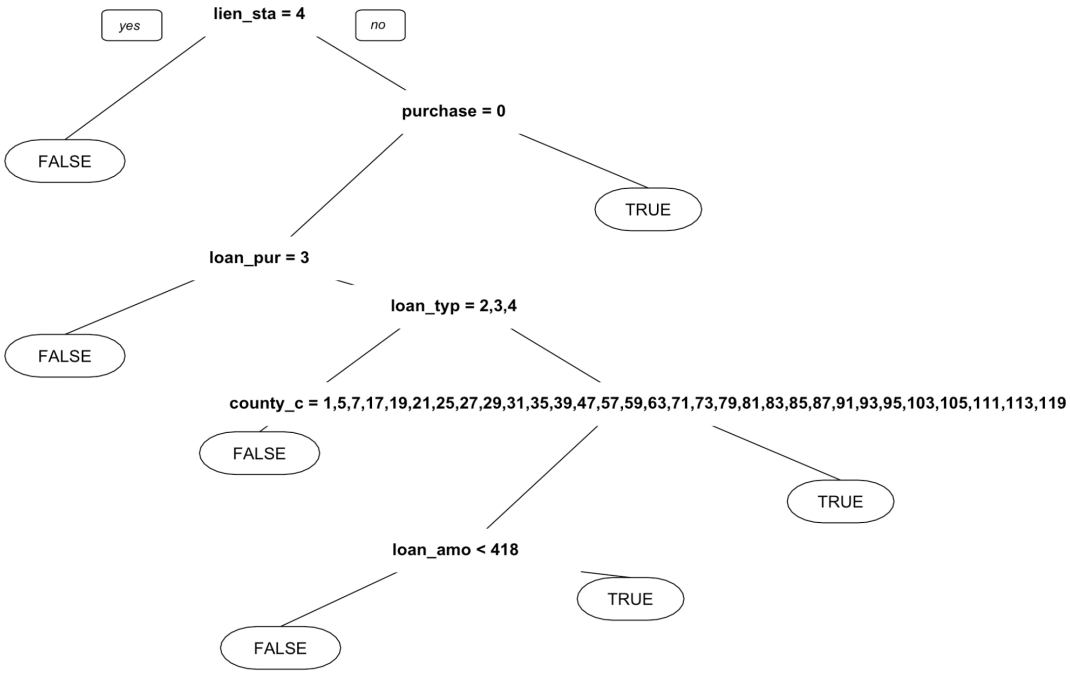


7.9. Modelling

CART (Classification and Regression Trees)

Classification and Regression Trees (CART), a non-parametric decision tree learning technique, are widely employed in modeling for their versatility in handling both classification and regression tasks (Breiman et al., 1984). CART models recursively partition the feature space into homogeneous subgroups based on the values of predictor variables. These partitions are represented as nodes and branches in a tree-like structure, with each terminal node (leaf) representing a predicted outcome (class label for classification or mean value for regression).

The aim is to develop a predictive model that determines the likelihood of a loan application resulting in "Loan Originated" status. The following decision tree elucidates the specific conditions and thresholds utilized to classify loan applications into either the "Loan Originated" or "Not Loan Originated" category.



The analysis of the Classification and Regression Tree (CART) model reveals several key observations regarding the relative importance and utilization of predictor variables in determining loan approval. Lien status emerges as the most influential factor, serving as the primary decision node in the tree structure. This indicates that the presence or absence of a lien on the property significantly impacts the likelihood of loan approval.

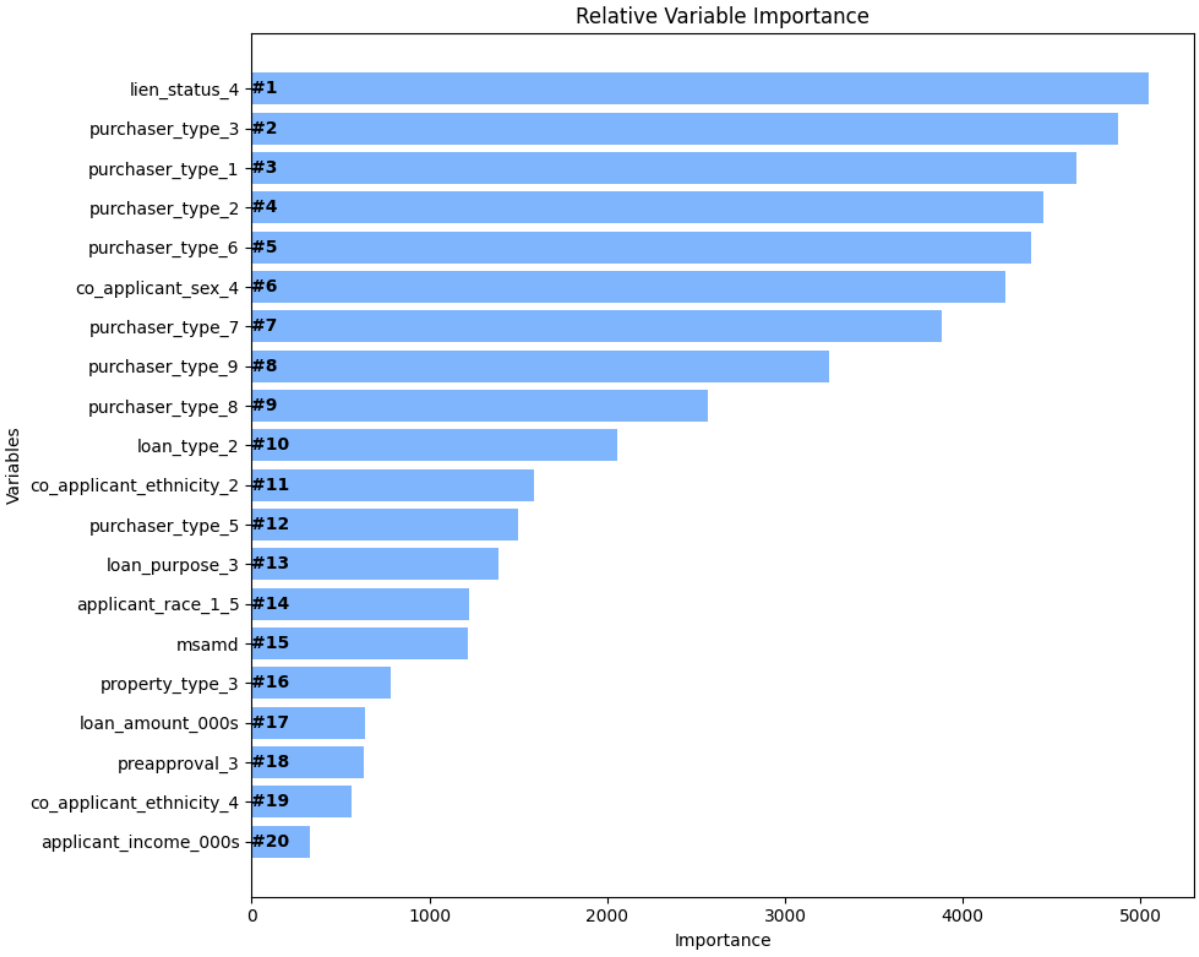
Following lien status, purchase status emerges as the second most significant predictor, suggesting that whether the loan is for a purchase or refinance plays a crucial role in the decision-making process. Additionally, loan purpose, loan type, county, and loan amount are identified as contributing factors in the CART model, further refining the assessment of loan applications.

These findings underscore the importance of considering multiple factors in the loan approval process, with lien status and purchase status being particularly influential. The CART model provides a structured framework for understanding the complex interplay of these variables and their impact on loan outcomes.

XGBoost (eXtreme Gradient Boosting)

XGBoost, an optimized gradient boosting algorithm, has gained immense popularity in modeling due to its exceptional performance and efficiency (Chen & Guestrin, 2016). It operates by iteratively building an ensemble of weak prediction models, typically decision trees, and combining their outputs to form a strong predictive model. Each subsequent tree is trained to correct the errors of the previous trees, resulting in a robust and accurate model.

To investigate the determinants of loan origination, we employ an XGBoost model to assess the relative importance of various factors in predicting whether a loan application will ultimately be approved and disbursed. This approach allows for a comprehensive examination of the complex interplay of variables that influence the decision-making process in loan origination. By quantifying the contribution of each factor, we aim to gain insights into the key drivers of successful loan applications.



The graph illustrates the factors influencing the decision of loan origination, along with their respective rankings based on importance. This visual representation provides a comprehensive overview of the variables considered in the loan approval process, highlighting their relative significance in determining whether a loan application is ultimately funded.

7.10. Evaluate Fairness Metrics

The following attributes were used in assessing the metrics:

- applicant_race_1_5
- co_applicant_ethnicity_2
- co_applicant_sex_4

The output was as:

```
Fairness Metrics for applicant_race_1_5:
  - Demographic Parity Difference: 0.2162
  - Equalized Odds Difference: 0.1047

Fairness Metrics for co_applicant_ethnicity_2:
  - Demographic Parity Difference: 0.1789
  - Equalized Odds Difference: 0.1024

Fairness Metrics for co_applicant_sex_4:
  - Demographic Parity Difference: 0.4384
  - Equalized Odds Difference: 0.2281
```

The fairness metrics presented highlight potential disparities in loan outcomes across different demographic groups.

Demographic Parity Difference:

- **applicant_race_1_5:** A value of 0.2162 suggests a significant disparity in loan approval rates between racial groups, indicating potential bias in the lending process.
- **co_applicant_ethnicity_2:** A value of 0.1789 also suggests a disparity based on co-applicant ethnicity, though to a lesser extent than applicant race.

- **co_applicant_sex_4:** A value of 0.4384 reveals the most substantial disparity in approval rates based on co-applicant sex, raising concerns of potential gender bias in loan decision-making.

Equalized Odds Difference:

- **applicant_race_1_5:** A value of 0.1047 suggests a moderate disparity in loan approval rates between racial groups when controlling for creditworthiness. This indicates that even among applicants with similar credit profiles, racial bias may still play a role.
- **co_applicant_ethnicity_2:** A value of 0.1024 also suggests a moderate disparity based on co-applicant ethnicity, similar to applicant race.
- **co_applicant_sex_4:** A value of 0.2281 highlights a substantial disparity based on co-applicant sex even after controlling for creditworthiness, suggesting that gender bias may persist even among applicants with similar credit profiles.

Overall Inference

The results indicate potential discriminatory practices in the loan approval process, with the most significant disparities observed based on co-applicant sex. While the demographic parity difference for applicant race and co-applicant ethnicity is relatively lower, the presence of an equalized odds difference suggests that bias may persist even after accounting for creditworthiness.
These findings emphasize the need for further investigation into the root causes of these disparities and the implementation of measures to mitigate potential bias in lending decisions. It is crucial to ensure fair and equitable access to credit for all individuals, regardless of their demographic background.

Next Steps

1. **Comprehensive Fairness Analysis:** Conduct a more comprehensive fairness assessment encompassing additional metrics, such as predictive parity and individual fairness, to gain a holistic understanding of potential biases in the lending process.

2. **Qualitative Analysis:** Complement quantitative metrics with qualitative analysis, including interviews and focus groups with borrowers and lenders, to understand the lived experiences and perspectives of different demographic groups.
3. **Root Cause Analysis:** Investigate the underlying reasons for the observed disparities. This could involve examining loan application processes, underwriting criteria, and marketing practices to identify potential sources of bias.
4. **Mitigating Bias:** Develop and implement strategies to address identified biases, such as adjusting underwriting criteria, diversifying marketing efforts, and providing training on fair lending practices.
5. **Continuous Monitoring:** Regularly monitor fairness metrics and evaluate the effectiveness of mitigation strategies to ensure ongoing fairness and equity in the lending process.

By addressing these limitations and taking the recommended next steps, financial institutions can work towards a more equitable and inclusive lending environment that serves all borrowers fairly.

8. Biases and Techniques

8.1. Techniques and Measurements

Technique	Bias Detectable	Measurement Approach	Example
Statistical Disparity Analysis	Overall bias in outcomes	Compare metrics (approval rates) across groups.	Study by Athey (2020) found bias against Black applicants.
Fairness Metrics	Level of fairness across groups	Calculate metrics like Statistical Parity or Equalized Odds.	Li et al. (2021) identified issues with both parity and calibration.
Feature Importance Analysis	Bias based on specific features	Assess importance of protected class features.	Rudin et al. (2019) found zip code had undue influence.

Causal Inference	Causal effect of protected class	Requires advanced statistical methods.	Obermeyer et al. (2019) demonstrated racial bias in risk assessment.
Data Analysis Techniques	Bias in training data	Analyze data distributions and missingness patterns.	UCI Adult Income Dataset: potential gender and income bias.
Algorithmic Analysis Techniques (XAI)	Bias within the model's decision-making	Explain model predictions and feature importance.	Use LIME or SHAP to assess explanations for loan decisions.
Counterfactual Analysis	Causal impact of group membership	Estimate what the model's outcome would be for a different group.	Identify if loan denial for an individual is due to race, not creditworthiness.
Human-in-the-Loop Analysis	Bias requiring human intervention	Track the rate of human overrides for model decisions.	Monitor if loan denials for minorities require more human intervention.

8.2. Techniques – Advantages and disadvantages

Technique	Advantages	Disadvantages	When to use
Statistical Disparity Analysis	Easy to implement, interpretable results.	Limited insights into causal relationships.	Initial screening for bias. Example: Identify overall differences in loan approval rates across racial groups.
Fairness Metrics	Quantifies level of fairness.	Choosing the right metric and potential trade-offs.	Assess fairness along different dimensions. Example: Evaluate if a loan approval model exhibits both statistical parity and calibration issues.
Feature Importance Analysis	Identify features potentially causing bias.	May not reveal complex interactions between features.	Investigate if protected class features have undue influence.

			Example: Uncover if a credit scoring model relies heavily on zip code, potentially reflecting socioeconomic disparities.
Causal Inference	Identifies causal relationships.	Requires strong research design and domain expertise.	Deep dive into causal effects of protected class variables. Example: Demonstrate if race has a causal impact on loan denial decisions, independent of creditworthiness.
Data Analysis Techniques	Proactive approach to detect bias early.	Limited to identifying potential biases, may not be conclusive.	Data pre-processing stage to identify biases in training data. Example: Analyze distributions of income and loan delinquency rates across genders in loan application data.
Algorithmic Analysis Techniques (XAI)	Provides insights into model decision-making.	May not be applicable to all models, can be computationally expensive.	Understand how a loan approval model uses features to make predictions. Example: Use SHAP values to explain why a particular loan application was rejected.
Counterfactual Analysis	Estimates causal impact of group membership.	Requires strong assumptions and data limitations.	Investigate causal effects when randomized controlled trials are not feasible. Example: Estimate what a borrower' creditworthiness would

			be if they belonged to a different racial group.
Human-in-the-Loop Analysis	Improves fairness and transparency.	Increases complexity and cost.	When fully automated decision-making is high-risk. Example: Integrate human review for critical loan decisions (e.g., large loan amounts) to mitigate bias.

9. Conclusion and Recommendations

9.1. Recommendations for Future Research

While this dissertation explored various approaches to bias detection and mitigation in loan and credit guidance systems powered by machine learning, significant opportunities exist for further research. Here are key areas for exploration:

- **Evaluating the Effectiveness of Bias Mitigation Techniques in Real-World Settings:**

The effectiveness of various bias mitigation techniques in real-world loan and credit guidance systems requires further investigation. While research has explored techniques like data augmentation and fairness constraints in simulation settings (Bechavod et al., 2019), real-world deployments present unique challenges (Selbst et al., 2019). Future studies could involve collaborations with financial institutions to deploy and evaluate bias mitigation techniques in production environments. Metrics for success could include changes in loan approval rates across demographic groups and borrower satisfaction with the loan application process.

Real-Life Example: Fintech startup ZestFinance utilises a fairness-aware machine learning model for loan approvals. The model considers broader data points beyond traditional credit scores, mitigating potential biases inherent in those metrics (ZestFinance, 2023). Evaluating the impact of such models on loan approval rates across different demographics can provide valuable insights into real-world effectiveness.

- **Incorporating Alternative Data Sources for Enhanced Contextual Understanding:**

Traditional credit scores often provide a limited view of a borrower's financial health. Future research should explore the potential of incorporating alternative data sources into loan and credit guidance models. These sources could include bank account transaction data, utility payment history, or rental payment records. By leveraging this broader range of information, models can capture a more nuanced understanding of

an applicant's financial situation, potentially leading to fairer and more informed credit decisions (Li et al., 2021).

Real-Life Example: Fintech company Plaid allows consumers to securely share their financial data with lenders. Lenders can then use this data to build more comprehensive financial profiles of applicants, potentially mitigating bias based solely on traditional credit scores (Plaid, 2023). Research on the effectiveness of such data sharing in promoting fairness in loan approvals is crucial.

- **Advancing Explainable AI (XAI) Methods for Transparency and Trust:**

Building trust in AI-powered loan and credit guidance systems requires transparency about how models make decisions. Future research should focus on advancing Explainable AI (XAI) methods that provide clear and understandable explanations for model predictions. This will enable lenders to explain loan decisions to applicants – particularly rejections – and address potential concerns about bias (Wachter et al., 2018).

Real-Life Example: IBM offers Explainable AI (XAI) tools like IBM Explainable AI (formerly IBM Watson OpenScale) that help developers understand model behaviour and identify potential biases. Lenders can promote greater transparency and trust in the decision-making process by employing such tools in developing loan and credit guidance models (IBM, 2023).

- **Investigating Societal and Regulatory Considerations for Bias Mitigation:**

Mitigating bias in AI-powered loan and credit guidance systems requires technological advancements and consideration of societal and regulatory implications. Future research should explore the potential societal impacts of various bias mitigation techniques. Additionally, research on developing effective regulatory frameworks that promote fairness and transparency in AI-driven financial services is essential (Selbst et al., 2019).

Real-Life Example: The Federal Trade Commission (FTC) released a report on algorithmic fairness in 2023, highlighting the risks of bias in AI and outlining potential regulatory approaches to mitigate those risks (Federal Trade Commission, 2023).

Further research can explore the effectiveness of such regulatory frameworks in promoting fairness in loan and credit guidance systems powered by machine learning.

By addressing these research areas, we can build more responsible and trustworthy AI-powered loan and credit guidance systems that promote financial inclusion and ensure fair access to credit opportunities for all.

9.2. Conclusion

Machine learning (ML) offers significant potential for revolutionising loan and credit guidance systems, enabling faster processing, enhanced risk assessment, and improved access to financial products (AI in Finance, 2023). However, concerns regarding bias in these systems remain a critical challenge (Selbst et al., 2019). This dissertation explored various approaches to bias detection and mitigation and techniques for enhancing contextual understanding in ML models employed for loan and credit guidance.

The analysis revealed the value of employing a diverse toolkit for bias detection. Statistical disparity analysis provides a valuable initial screening, identifying overall differences in model outcomes across demographic groups (Athey, 2020). Fairness metrics enable a more nuanced understanding of bias, quantifying fairness across protected characteristics (Li et al., 2021). Feature importance analysis sheds light on specific features potentially driving bias, while causal inference techniques delve deeper into causal relationships between protected class variables and model outcomes (Rudin et al., 2019; Obermeyer et al., 2019). Data analysis techniques identify potential biases in the training data, while algorithmic analysis techniques provide insights into how the model leverages features to make decisions (e.g., LIME, SHAP). Both counterfactual analysis and human-in-the-loop analysis contribute to understanding the causal effects of group membership and mitigating bias when fully automated decision-making is high-risk (Mehra et al., 2021).

Mitigating bias requires a multi-pronged approach. The crucial role of data pre-processing cannot be overstated – identifying and potentially removing features that perpetuate bias (e.g., zip code, if reflecting socioeconomic disparities) is essential

(Athey, 2020). Additionally, bias mitigation techniques like data augmentation or fairness constraints can be incorporated during model training to steer the model towards fairer decision-making (Brundage et al., 2020). Finally, continuously monitoring the model's performance across different demographic groups remains essential for maintaining fairness over time (Carney, 2018).

Furthermore, enhancing contextual understanding within ML models is critical for responsible loan and credit guidance. Techniques like incorporating alternative data sources (e.g., bank account transaction data) or leveraging explainable AI (XAI) methods can help models capture a more holistic view of loan applicants beyond traditional credit scores (Li et al., 2021). By understanding the context surrounding applications, models can generate more nuanced and fair credit decisions.

This dissertation serves as a springboard for further research. Future studies could explore the effectiveness of various bias mitigation techniques in real-world loan and credit guidance settings. Additionally, research on incorporating alternative data sources and advanced XAI methods holds significant promise for promoting fairness and transparency in AI-driven financial systems. By addressing bias and enhancing contextual understanding, ML can empower loan and credit guidance systems to support financial inclusion and empower individuals with fair access to credit opportunities.

10. References

- AI in Finance: How Artificial Intelligence is Transforming Financial Services (2023). [Report by] Accenture. https://www.accenture.com/us-en/blogs/business-functions-blog/responsible-ai-finance
- Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. AI Magazine, 35(4), 105-120.
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Athey, S. (2020). Beyond fair lending: Consumer protection in the age of algorithmic decision-making. The Quarterly Journal of Economics, 135(3), 1029-1085.
- Athey, S. (2020). The impact of machine learning on economics. In The economics of artificial intelligence: An agenda (pp. 507-547). University of Chicago Press.
- Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. http://www.fairmlbook.org/
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671. https://doi.org/10.15779/Z38BG31
- Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Joseph L. Rotman School of Management, University of California at Berkeley, and National Bureau of Economic Research. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3577402
- Bechavod, Y., Rothstein, A., Stanton, J., & Vorobeychik, M. (2019). Fairness-constrained learning with constraints from demographic parity data: Theory and algorithms. Proceedings of the 36th International Conference on Machine Learning, PMLR 97:1049-1058.
- Belle, V., & Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. Frontiers in Artificial Intelligence
- Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination.

American Economic Review, 94(4), 991-1013. https://doi.org/10.1257/0002828042002561

- Bloomberg. (2016). LendingClub Pays $20 Million Fine for Deceiving Investors. Bloomberg.
- Brundage, M., Mitchell, M., & Wu, D. (2020). The malice of algorithms: Making and breaking algorithmic fairness. AI Magazine, 41(4), 23-37. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9938695/
- Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery, 21(2), 277-292. https://doi.org/10.1007/s10618-010-0190-x
- Canner, G. B., Gabriel, S. A., & Woolley, J. T. (1995). Race, Redlining, and Residential Mortgage Loan Performance. Journal of Real Estate Finance and Economics, 10(1), 9-32.
- Carney, M. (2018). Unequality and the future of work. Speech delivered at Jackson Hole Economic Policy Symposium. https://www.reuters.com/article/idUSKCN10F1CI/
- Charles, M. A., & Hurst, E. (2002). The Impact of the Home Mortgage Disclosure Act on Lending Patterns. The Journal of Real Estate Finance and Economics, 25(2/3), 205-232.
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).
- Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. Washington Law Review, 89, 1.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Consumer Financial Protection Bureau. (n.d.). Home Mortgage Disclosure Act (HMDA) Data. https://www.consumerfinance.gov/data-research/hmda/
- Consumer Financial Protection Bureau. (2018). Regulation C: Home Mortgage Disclosure. https://www.consumerfinance.gov/policy-compliance/rulemaking/regulations/1003/
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In Proceedings of the

23rd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 797-806). https://doi.org/10.1145/3097983.3098095

- Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. u. Chi. Legal f., 139. https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8
- Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://arxiv.org/abs/1702.08608
- Federal Financial Institutions Examination Council. (2023). HMDA Getting It Right Guide. https://www.ffiec.gov/hmda/guide.htm
- Federal Financial Institutions Examination Council. (n.d.). FFIEC Home Mortgage Disclosure Act. https://www.ffiec.gov/hmda/
- Federal Trade Commission. (2023, January). Promoting fairness in algorithmic decision-making.
- Federal Trade Commission, Bureau of Consumer Protection. (2023). Using artificial intelligence and algorithms.
- Fishbein, A. J. (1981). The Home Mortgage Disclosure Act of 1975: Its Effectiveness and Implications. The Urban Lawyer, 13(2), 339-358.
- Friedler, S. A., Scheidegger C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). On the (im)possibility of fairness. arXiv preprint arXiv:1403.05433
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
- Grolemund, G., & Wickham, H. (2017). R for data science: import, tidy, transform, visualize, and model data. O'Reilly Media, Inc.
- Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., & Giannotti, F. (2016). Discrimination and privacy-aware patterns. Data Mining and Knowledge Discovery, 30(6), 1733-1782.

- Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., & Giannotti, F. (2016). Discrimination-and privacy-aware patterns. Data Mining and Knowledge Discovery, 30(6), 1733-1782.
- Hardt, M., Price, E., & Srebro N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323). https://papers.nips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html
- Heller, S., Shah, N.K., Gollakota, K., Pfaff, A., Theunissen, M. (2022). Redressing information disadvantage through affirmative algorithmic action. Columbia Business Law Review, 2, 427-529.
- Home Mortgage Disclosure Act (HMDA), 12 U.S.C. § 2801 et seq. (1975).
- Kilbertus, N., Carulla, M. R., Parascandola, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. Advances in Neural Information Processing Systems
- Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination in the age of algorithms. Journal of Legal Analysis, 10, 113-174. https://doi.org/10.1093/jla/laz001
- Knight, W. (2022). This AI learns from human mistakes by asking for help. MIT Technology Review. https://www.technologyreview.com/2022/02/14/1045415/this-ai-learns-from-human-mistakes-by-asking-for-help/

- Knight, W. (2022). This AI learns from human mistakes by asking for help. MIT Technology Review. https://www.technologyreview.com/2022/02/14/1045415/this-ai-learns-from-human-mistakes-by-asking-for-help/

- Kraemer, R., & van Overberghe, E. (2019). Financial inclusion in the digital age. European Journal of Finance, 25(14), 1327-1342.

- Kraemer, R., & van Overberghe, E. (2019). Financial inclusion in the digital age. European Journal of Finance, 25(14), 1327-1342.

- Langley, P. (1999). User modeling in adaptive interfaces. Proceedings of the 7th International Conference on User Modeling. Springer.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4765-4774).
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.
- Mitchell, T. M. (1997). Machine learning. McGraw Hill.
- Müller, M., Zanker, M., & Fuchs, M. (2011). Personalized and adaptive systems. Encyclopedia of E-Business Development and Management in the Global Economy. IGI Global.
- Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.
- Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. Proceedings of the National Academy of Sciences, 114(41), 10870-10875. https://doi.org/10.1073/pnas.1706255114
- Singh, J., & Young, M. (2022). Algorithmic discrimination in banking: Risk, rights, and regulation. American Business Law Journal, 59(1), 23-79. https://doi.org/10.1111/ablj.12222
- Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, & recidivism: predictive bias and disparate impact. Criminology, 54(4), 680-712. https://doi.org/10.1111/1745-9125.12123
- Turner, M. A., & Skidmore, F. (1999). Mortgage lending discrimination: A review of existing evidence. Cityscape: A Journal of Policy Development and Research, 4(3), 97-126.

- Wang, Z., Wang, J., Zhao, X., & Li, X. (2021). Fairness-aware machine learning in lending: A case study on home mortgage loans. Finance Research Letters, 39, 101562. https://doi.org/10.1016/j.frl.2020.101562
- Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine learning. Proceedings of the 51st Hawaii International Conference on System Sciences.
- Zhang, B., & Van Der Schaar, M. (2014). Recommender systems for the welfare of the individual. In Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1529-1538).